Look Further to Recognize Better: Learning Shared Topics and Category-Specific Dictionaries for Open-Ended 3D Object Recognition

Mohades Kasaei, Hamidreza

Published in:
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)

DOI:
10.1109/IROS40897.2019.8967823

2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Look Further to Recognize Better: Learning Shared Topics and Category-Specific Dictionaries for Open-Ended 3D Object Recognition

S. Hamidreza Kasaei

***Abstract*— Service robots are expected to operate effectively in human-centric environments for long periods of time. In such realistic scenarios, fine-grained object categorization is as important as basic-level object categorization. We tackle this problem by proposing an open-ended object recognition approach which concurrently learns both the object categories and the local features for encoding objects. In this work, each object is represented using a set of *general latent visual topics* and *category-specific dictionaries*. The general topics encode the common patterns of all categories, while the category-specific dictionary describes the content of each category in details. The proposed approach discovers both sets of general and specific representations in an unsupervised fashion and updates them incrementally using new object views. Experimental results show that our approach yields significant improvements over the previous state-of-the-art approaches concerning scalability and object classification performance. Moreover, our approach demonstrates the capability of learning from very few training examples in a real-world setting. Regarding computation time, the best result was obtained with a Bag-of-Words method followed by a variant of the Latent Dirichlet Allocation approach.**

I. Introduction

Nowadays service robots are leaving the structured and completely known environments and entering human-centric settings. For these robots, object perception is a challenging task due to the high demand for accurate and real-time response under changing and unpredictable environmental conditions. Although many problems have already been understood and solved successfully, many challenges still remain. Open-ended object recognition is one of these challenges waiting for many improvements. Cognitive science revealed that humans learn to recognize object categories ceaselessly over time. This ability allows adapting to new environments, by enhancing their knowledge from the accumulation of experiences and the conceptualization of new object categories. Inspired by this, we approach object category learning and recognition from a long-term perspective and with emphasis on open-endedness. In this paper, *open-ended* implies that the set of object categories to be learned is not known in advance, and the training instances are extracted from online experiences of a robot, and become gradually available over time, rather than being completely available at the beginning of the learning process.

In such a complex, dynamic and realistic setting, no matter how extensive the training data used for batch learning, a robot might always face a new object. Therefore, apart from batch learning, the robot should be able to learn new object categories from very few training examples on-site supported by human-in-the-loop feedback. Moreover, the robot may frequently face a new object that visually can be either *not similar* (*basic-level*) or *very similar* (*fine-grained*) to the previously learned object categories (see Fig. 1). This poses a significant challenge in situations when the robot needs to recognize visually similar categories for which only a few examples are available. In such situations, *object representation* plays a central role because the output of this module is used for learning as well as recognition. In this paper, we propose a new method to characterize an object based on a *common set of latent visual topics*, which is used to describe the content of all categories (basic-level), and a set of *category-specific dictionaries* for highlighting the small diversity within the different categories (fine-grained). The representation of the given object is finally obtained by concatenating the generic and specific representations. To the best of our knowledge, there is no other approach that can jointly learn a set of generic and category-specific features for encoding 3D object categories in an open-ended manner.

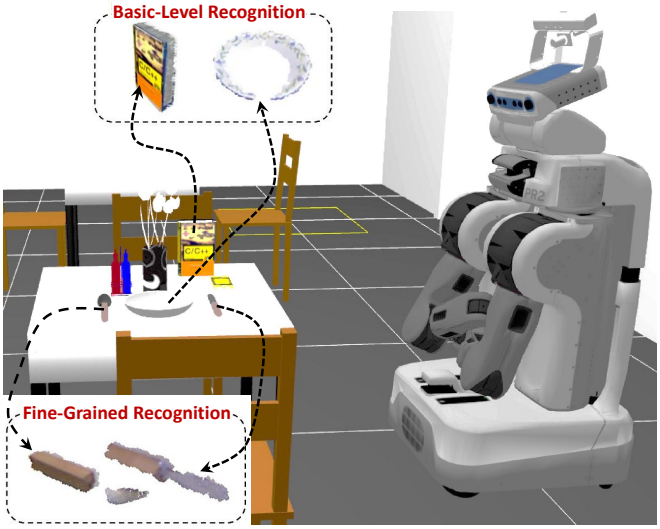


Fig. 1. An illustrative example of basic-level object categorization vs. fine-grained object categorization. PR2 robot is looking at some objects on the table. In such scenarios, distinguishing *Spoon* from *Knife* (fine-grained) is harder than distinguishing *Book* from *Plate* (basic-level) due to subtle differences in particular regions.

II. Related Work

In the last decade, various research groups have made substantial progress towards the development of learning approaches which support online and incremental object

Department of Artificial Intelligence, University of Groningen, PO Box 407, 9700 AK, Groningen, The Netherlands. Email: hamidreza.kasaei@rug.nl

We are thankful to Prof. L. Seabra Lopes, Prof. A. Tomé, and Prof. M. Wiering who provided expertise that greatly assisted the research.

category learning [1][2]. In recent studies on object recognition, much attention has been given to deep Convolutional Neural Networks (CNNs). It is now clear that if in a scenario, we have *a fixed set of object categories* and *a massive number of examples per category that are sufficiently similar to the test images*, CNN-based approaches yield good results, notable recent works include [3][4]. In open-ended scenarios, these assumptions are not satisfied, and the robot needs to learn new concepts on-site using very few training examples. While deep learning is a very powerful and useful tool, there are several limitations to apply CNNs in open-ended domains. In general, CNN approaches are incremental by nature but not open-ended, since the inclusion of new categories enforces a restructuring in the topology of the network. Furthermore, training a CNN-based approach requires long training times and training with a few examples per category poses a challenge for these methods. In contrast, *open-ended learning* [5][6] allows for concurrent learning and recognition. Our approach falls into this category.

In the case of object representation, most of the recent approaches use either neural networks [7][4] or hand-crafted features [8]. These approaches may not be the best option for such domains since the object representation procedure is a built-in component of the system. Oliveira et al. [9] tackle this problem by proposing an approach for concurrent learning of visual codebooks and object categories in open-ended domains. Unlike our approach, they completely discard information related to the co-occurrence of local object features. Moreover, they did not consider fine-grained categorization. Existing approaches for fine-grained categorization heavily rely on accurate object parts/features annotations during the training phase. Such requirements prevent the wide usage of these methods. However, some works only use class labels and do not need exhaustive annotations. Geo et al. [10] proposed a Bag of Words (BoW) approach for fine-grained image categorization. This work is similar to ours since they represent objects using generic and specific representations. However there are some differences: their codebooks are constructed offline; therefore, the representativeness of the training set becomes critical to the performance of the system. Moreover, this approach is impractical for an open-ended domain (i.e., large number of categories) since the size of object representation is linearly dependent on the number of known categories. Zhang et al. [11] proposed a novel fine-grained image categorization system. Similar to our work, they only used class labels during the training phase. Unlike our approach, in [10][11] the representations of known categories do not change after the training stage. Moreover, they completely discarded co-occurrence (structural) information. This limitation might lead to non-discriminative object representations and, as a consequence, to poor object recognition performance.

Several types of research have been performed to assess the added-value of structural information. Canini et al. [12] extended an online version of Latent Dirichlet Allocation (LDA) and proposed an incremental Gibbs sampler for LDA (here referred to as I-LDA). In online-LDA and I-LDA, the number of categories is fixed, while in our approach the number of categories is growing. Kasaei et al. [5] proposed an open-ended object category learning approach just by learning specific topics per category, while our approach does not only learn a set of general topics for basic-level categorization, but also learn a category-specific dictionary for fine-grained categorization.

III. Learning Generic and Category-Specific Representations

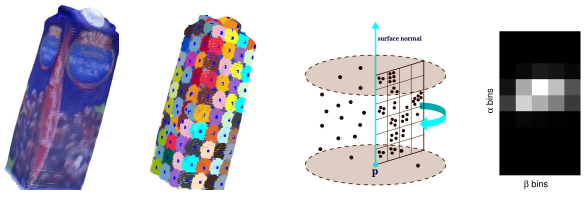
We organized our approach in three main phases: (*i*) feature extraction; (*ii*) learning generic representations; and (*iii*) learning category-specific representations. In the following we describe each of these phases in detail.

A. Feature Extraction

In this work, we first represent a 3D object by a set of local shape features called spin-image [13]. The reason why we use spin-image rather than other 3D feature descriptors is that the spin-image is a pose-invariant feature, and therefore suitable for 3D perception in autonomous robots. Another advantage of the spin-image is that it only requires a surface normal - rather than a full reference frame - to compute the feature. As depicted in Fig. 2, the process of local feature extraction consists of two main phases: extraction of key points and computation of spin-images. For efficiency reasons, the number of key points in an object should be much smaller than the total number of points. Therefore, the object is first voxelized and then, the nearest neighbor point to each voxel center is selected as a key point (Fig. 2 *a* and *b*). Afterwards, the spin-image descriptor is used to encode the surrounding shape in each keypoint using the original point cloud. A spin-image is a local shape histogram, which is obtained by spinning a 2D histogram around the key point's surface normal (see Fig. 2 *c* and *d*). Therefore, each object view is described by a set of spin-images, $\mathbf{O} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$, where N is the number of key points. The obtained representation is then used as input to both generic and category-specific object representation processes.

B. Generic Representation

Our generic object representation approach requires a dictionary with V visual words. Usually, the dictionary is created via off-line clustering of training data. In this work, we create a pool of objects using 75% of the training data. Previously, we showed how a robot could create a pool of



(a) *(b)* *(c)* *(d)*

Fig. 2. Local feature extraction: *(a)* point cloud of a juice-box *(b)* key point selection (blue points); *(c)* a schematic of how the spin-image spins a 2D histogram around the key point's surface normal; *(d)* the computed spin-image for a sample keypoint.

objects by exploring the environment [5]. To construct a pool of features, spin-images are computed for the selected points extracted from the pool of objects. Finally, the dictionary is constructed by clustering the features using the k-means algorithm [14]. The centers of V computed clusters are defined as the visual words, $\mathbf{w}_i : i \in \{1, \ \ldots, \ V\}$.

We then modify the incremental LDA approach [12] to be suitable for this study by releasing the batch learning phase. Particularly, we have tried to get structural semantic features from low-level feature co-occurrences by discovering a set of topics using an incremental LDA model. The basic idea is to represent a given object as a histogram of topics (i.e., latent variables), where a distribution over visual words characterizes each topic. In this method, an object is first represented as a set of visual words, $\{\mathbf{s}_1, \ \ldots, \ \mathbf{s}_N\} \rightarrow \{w_1, w_2, ..., w_N\}$, where each entry represents the index of one of the V words of the dictionary. Next, the object should be described as a set of topics, $\{w_1, w_2, ..., w_N\} \rightarrow \{z_1, z_2, ..., z_N\}$, where each element of the topic set represents the index of one of the K topics. Towards this goal, the probability of topic $\mathbf{z}_j$ being assigned to a word $\mathbf{w}_i$, given all other topics assigned to all other words is estimated by the fast collapsed Gibbs sampler. After the sampling procedure, the word-topic matrix can be estimated as follows:

$$\phi_{w_i,k} = \frac{n_{w_i,k} + \beta}{n_k + |\mathbf{V}| \times \beta}, \tag{1}$$

where β is Dirichlet prior hyper-parameter that affect the sparsity of distributions, $n_{\mathbf{w}_i,k}$ shows the number of times visual word $\mathbf{w}_i$ was assigned to topic k and n_k is the number of times a word was assigned to topic k. ϕ is a $K \times V$ matrix, which represents words-probability for each topic, K is the number of topics, and $\phi_{w_i,k} = p(\mathbf{w}_i|\mathbf{z}_k)$. As it is shown in Algorithm 1, we also need $n_{o,k}$ counter through the sampling procedure, which shows the number of times topic k is assigned to some visual words of object $\mathbf{O}$. After inferring the word-topic matrix, we generate a set of K spin-image like topics:

$$\mathbf{z}_k = \sum_{i=1}^{V} p(\mathbf{w}_i|\mathbf{z}_k) \times \mathbf{w}_i \qquad k \in \{1, \ldots, K\}. \tag{2}$$

Each topic is then normalized to provide further robustness to depth variations of the object in the scene. In this way, each topic is generated by combining all visual words. It is worth mentioning that the process of topic learning does not require an explicit distance metric. The proposed procedure is summarized in Algorithm 1. At some points, a user may instruct the robot to update topics. In this case, the robot retrieves the representation of all the instances of all categories and updates dictionaries as well as the parameters of the model including n_k and $n_{w,k}$ incrementally (i.e., unsupervised learning) using Gibbs sampling. Figure 3 shows a sample set of learned topics. For representing a given object using the learned topics, each local feature of the object is approximated by its nearest topic. Then the object is represented as a histogram of occurrences of topics $\mathbf{h}^t = [h_1, \ h_2, \ \ldots, \ h_K]$ where the i^{th} element of $\mathbf{h}^t$ is the count of the number of spin-images assigned to a topic, $\mathbf{z}_i$.

Algorithm 1 Generic representation learning

Input:
$W \leftarrow$ dictionary of visual words
$V \leftarrow$ size of dictionary
$K \leftarrow$ number of topics
$G \leftarrow$ number of Gibbs sampling iterations
$\alpha, \beta \leftarrow$ Dirichlet prior hyper-parameters
object_view o is given as a sequence of N visual words, $\{w_1, \ldots, w_N\}$

Input/Output:
count_matrices $n_{o,k}$, $n_{w,k}$, n_k

Initialization:
for all visual words $w_i : i \in \{1, \ldots, N\}$ **do**
  $k \leftarrow z_i$ ▷ sample a topic index randomly
  increment $n_{o,k}$, $n_{w,k}$, n_k
end for

Inference:
for $it \in \{1, \ldots, G\}$ **do** ▷ Gibbs sampling
  for all visual words $w_i : i \in \{1, \ldots, N\}$ **do**
    $k \leftarrow z_i$ ▷ sample a topic index
    decrement $n_{o,k}$, $n_{w,k}$, n_k
    for $k \in \{1, \ldots, K\}$ **do**
      $p[k] \leftarrow p[k-1] + (n_{o,k} + \alpha) \times \frac{n_{w,k}+\beta}{n_k + V\beta}$
    end for
    $u \leftarrow$ draw from uniform [0, 1]
    for $k \in \{1, \ldots, K\}$ **do** ▷ sample the topic index
      if $u < p[k]$ / $p[K]$ **then**
        $z_i \leftarrow k$
        stop
      end if
    end for
    increment $n_{o,k}$, $n_{w,k}$, n_k ▷ update the counters
  end for
end for

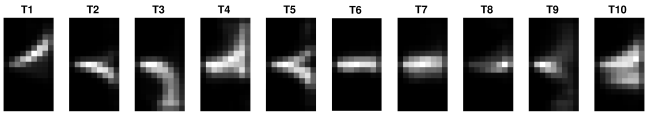


Fig. 3. A sample set of latent topics: spin-images compiled with IW = 8 bins and SL = 0.09m.

C. Category-Specific Representation

While the generic representation works well for basic-level classification, it does not work for fine-grained categorization. The underlying reason is that most features from fine-grained categories are similar. Therefore, these categories would share lots of similar topics, and the proportion of discriminative topics would be minor. Therefore, it is desirable to develop an approach for learning a set of category-specific features from a few examples. We have tackled this problem by learning a category-specific dictionary for each category independently. This is a challenging task since the robot does not know in advance which object categories it will have to learn, which observations will be available, and when they will be available to support learning. In our current setup, a new instance of a specific category is stored in the robot's memory in the following situations:

- When the teacher for the first time *teaches* a particular category, through a *teach* action, a new category is created and initialized with the set of object views.
- In the case of *correct* actions, the local features of the target object view are added to the category.

Assume at time t_0 (i.e., first teaching action) a dictionary is learned for category c, denoted as $V^c_{t_0}$, which represents

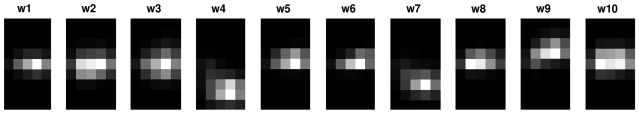


Fig. 4. A sample set of visual words of a category-specific dictionary: spin-images compiled with IW = 4 bins and SL = 0.09m.

the distribution of 3D shape features observed up to time t_0. Later at time t_1, a new training instance, which is represented as a set of spin-images, is taught by a teacher to category c (i.e., supervised learning). The teaching instruction trigs the robot to retrieve the current dictionary of the category as well as the representation of the new object view and updates the relevant dictionary using an incremental K-means algorithm [14] (i.e. unsupervised learning). Such category-specific dictionary would highlight the differences of objects from different categories, and as a consequence improves the object recognition performance.

Similar to the generic representation, a given object is then represented as a histogram of occurrences of visual words, $\mathbf{h}^c = [h_1,\ h_2,\ \ldots,\ h_{V^c}]$, where V^c is the size of category-specific dictionary. The obtained histograms, $\mathbf{h}^t$ and $\mathbf{h}^c$, are then concatenated to form a single representation for the given object. Figure 4 shows ten sample visual words from a *Mug* category-specific dictionary. It is worth mentioning when dictionaries or topics are updated, the representations of known instances must be updated. At the moment, we do this by storing the features of each object view and using them to recompute the representations.

IV. Object category learning and recognition

Concerning category formation, we use the instance-based learning and recognition (IBL) approach which considers category learning as a process of learning about the instances of the category, i.e., a category is represented simply by a set of known instances. IBL is a baseline approach to evaluate object representations. An advantage of the IBL approaches is that they can recognize objects using a very small number of experiments and the training phase is very fast. As discussed in the previous section, the *teach* and *correct* actions by the teacher lead the robot to create a new category or to modify an existing one (i.e., supervised learning). Whenever a new object is added to a category, the agent retrieves the current model of the category and updates the category model by storing the representation of new object views. In particular, our approach can be seen as a combination of a particular *object representation*, *similarity measure* and *classification rule*. Therefore, the choice of the similarity metric has an impact on the recognition performance.

With regards to the similarity measure, since the proposed object representation describes an object as a histogram, the dissimilarity between two histograms can be computed by different distance functions. We refer the reader to a comprehensive survey on distance/similarity measures provided by Cha [15]. After performing several cross-validation experiments, we concluded that two types of distance functions including Jensen-Shannon (JS) and chi-squared (χ^2) distances are suitable to estimate the similarity between two instances. Both functions are in the form of a bin-to-bin distance function. Although the practical results of χ^2 and JS are almost identical, χ^2 is computationally more efficient. Therefore, we use the χ^2 function to estimate the similarity of two instances. Mathematically, let P and Q $\in \mathbb{R}^{K+V^c}$ be the representation of two objects:

$$\chi^2(\mathrm{P},\mathrm{Q}) = \frac{1}{2}\sum_{i=1}^{K+V^c}\frac{(\mathrm{P}_i-\mathrm{Q}_i)^2}{(\mathrm{P}_i+\mathrm{Q}_i)}. \quad (3)$$

To assess the dissimilarity between a target object and stored instances of a certain category **C**, the target object should be first described by the general topics and the learned dictionary of the category **C**. Afterwards, the minimum distance between the target object and all stored instances of the category **C** is considered as the Object-Category-Distance (OCD). The target object is finally classified based on the minimum OCD.

V. Result and discussion

Three types of experiments were carried out to evaluate the proposed approach. In this section, we first explain our experimental setup and then discuss the obtained results.

A. Datasets and Baselines

The experimental evaluations were carried out on two 3D object datasets including the Restaurant Object dataset [6] and the Washington RGB-D Object dataset [16]. For the classical evaluation (i.e., 10-fold cross-validation), we mainly use the Restaurant Object Dataset since it has a small number of classes (10 categories) with significant intra-class variation that is suitable for performing extensive sets of experiments. The Washington RGB-D Object dataset [16] is used for open-ended evaluation. It consists of 250,000 views of 300 common household objects taken from multiple views and organized into 51 categories. We also report on a real-world demonstration using the Imperial College Domestic Environment Dataset [17]. This is a suitable dataset for this test since all scenes were captured under various clutter and contain objects with similar shapes (*lipton* vs. *softkings*) and objects with very different shapes (*oreo* vs. *elite*).

For comparison, we have selected four open-ended 3D object category learning and recognition approaches, including the RACE [1], BoW [18] based on the nearest neighbour classification rule, and Open-Ended LDA, which is a modified version of the standard smoothed LDA [19] and Local-LDA [5]. Moreover, we add another baseline, which is the proposed method without category-specific representation (here referred to as Generic Rep.).

B. Classical Evaluation using Restaurant Object Dataset

The proposed approach has a set of parameters including $\langle \alpha,\ \beta,$ G, VS, IW, SL, V, K and $\mathrm{V}^c\rangle$, that must be tuned to provide a good balance between recognition performance, memory usage, and processing speed. In this work, we assumed a symmetric Dirichlet prior for both α and β parameters. Therefore, a high α value means that each object is likely to contain a mixture of most of the topics, and

TABLE I

OBJECT RECOGNITION PERFORMANCE FOR DIFFERENT PARAMETERS USING RESTAURANT OBJECT DATASET.

Parameters	VS (m)		IW (bins)		SL (m)			V (generic words)					K (topics)			V^c (specific words)				
Values	0.02	0.04	4	8	0.07	0.08	0.09	50	60	70	80	90	70	80	90	50	60	70	80	90
Avg. Accuracy(%)	**85**	81	83	**83**	82	83	**83**	82	82	83	84	**84**	**84**	83	82	81	84	**84**	83	82

not a single specific topic. Likewise, a low β value means that a topic may contain a mixture of just a few of the words. We carried out several experiments and concluded that α and β should be set to 1.0 and 0.1 respectively. The number of Gibbs sampling iterations, G, is set to 50. A set of 900 experiments was performed for different values of the remaining six parameters. The voxel size (VS) parameter is related to the number of keypoints extracted from each object view. IW defines the size of the spin-image descriptor, which will be $(\mathrm{IW}+1)\times(2\times\mathrm{IW}+1)$ float (4 bytes). Support length (SL) determines the amount of space swept out by a spin-image. A summary of the experiments using Restaurant Object dataset [6] is reported in Table I. The parameter configuration that obtained the best *average accuracy* was selected as the default configuration (i.e., bold numbers). The accuracy of this configuration was 94%. A complete experiment (including both learning and recognition phases) on average took 286.50 seconds.

1) Comparative evaluation: Table II presents a summary of the obtained results. One important observation is that the overall performance of our approach is promising and the proposed representation is capable of providing a distinctive representation for the objects. Moreover, it was observed that the discriminative power of our approach was better than the other evaluated approaches which was 7 percentage points (p.p.) better than RACE, 5 p.p. better than BoW, 11 p.p. and 3 p.p. better than Open-Ended LDA (shared topics) and Local LDA (category-specific topics) respectively. Furthermore, to show the importance of considering the category-specific representation, we compared our approach with the Only Generic Rep. baseline. It was observed that the accuracy has been boosted by 4 p.p. when we use both generic and category-specific representations. The accuracy of object recognition based on variable size representation (i.e., RACE) was not as good as the other approaches. The Local LDA and BoW obtained an acceptable performance. In the case of experiment times, BoW achieves the best performance and RACE was the most computationally expensive approach. Overall, topic modelling approaches achieve a medium-level experiment time. The underlying reason is that there is a Gibbs sampling procedure in the topic modelling based approaches which takes time to accurately represent the desired distribution.

TABLE II

RECOGNITION PERFORMANCE

Approach	Accuracy	Time (s)
RACE[1]	0.87	1757
BoW[18]	0.89	**196**
LDA[19]	0.83	258
Local-LDA[5]	0.91	349
Generic Rep.	0.90	234
Our Work	**0.94**	287

C. Open-Ended Evaluation

An evaluation protocol for open-ended learning systems was proposed in [20]. The idea is to emulate the interactions of a robot with the surrounding environment over large periods of time. This protocol is based on a Test-then-Train scheme.We developed a *simulated teacher* to follow the protocol and autonomously interact with the agen. The idea is that for each newly taught category, the simulated teacher repeatedly picks unseen objects of the currently known categories from a dataset and presents them to the agent. It progressively estimates the recognition accuracy of the agent and, in case this accuracy exceeds a given threshold (τ = 0.67, meaning accuracy is at least twice the error rate), introduces an additional object category. This way, the agent is trained online, and at the same time, the accuracy of the system is continuously estimated. In case the agent can not reach the classification threshold after a certain number of iterations (i.e., 100 iterations), the simulated teacher can infer that the agent is no longer able to learn more categories and terminates the experiment. We assess our experimental results using the metrics that were recently introduced in [9], including: (*i*) the average number of learned categories at the end of an experiment (ALC), an indicator of *how much the system is capable of learning*; (*ii*) the number of question/correction iterations (QCI) required to learn those categories and the average number of stored instances per category (AIC), indicators of *how much memory does it take for learning?*; (*iii*) Global Classification Accuracy (GCA), an accuracy computed using all predictions in an experiment, and the Average Protocol Accuracy (APA), indicators of *how well the system learns*.

1) Results: A detailed summary of the obtained results is reported in Table III. In all approaches, as more categories are learned, the classification accuracy first decreases (performance degradation phase), and then starts going up again as more instances are introduced (recovery phase). Eventually, the agent reaches a break-point where it is no longer able to learn more categories. By comparing all approaches, it is visible that the agent learned (on average) more categories using our approach than with other state-of-the-art approaches. The agent with Local-LDA approach obtained acceptable scalability, while the scalability of the other approaches was very low and their performance drops aggressively when the number of categories increases. In particular, our approach on average learned around 11 categories more than Local-LDA and 18, 20 and 25 categories more than BoW, RACE and open-ended LDA approaches, respectively. It is also clear that the agent with our approach stored more instances per category (AIC) than the other approaches. This is expected since a higher number of categories known by the system tends to make the classification task more difficult. Although on average, BoW and Local-LDA achieved better performance than our approach, the difference is minor, and the discriminative power of those approaches is lower. This is expected since the BoW and Local-LDA learned

TABLE III

SUMMARY OF OPEN-ENDED EVALUATION.

Approaches	#QCI	ALC	AIC	GCA	APA
RACE[1]	382.10	19.90	8.88	0.67	0.78
BoW[18]	411.80	21.80	8.20	0.71	0.82
LDA[19]	262.60	14.40	9.14	0.66	0.80
Local-LDA[5]	613.00	28.50	9.08	0.71	0.80
Our work	1344.40	**39.80**	13.12	0.70	0.79

fewer categories, and it is easier to get better classification performance in fewer categories.

D. System Demonstration

In this demonstration, the system initially had no prior knowledge, and all objects are recognized as *Unknown*. Later, a user interacts with the system and teaches all object categories including *amita*, *colgate*, *lipton*, *elite*, *oreo* and *softkings* to the system using the objects extracted from a scene captured from the blue camera pose as shown in Fig. 5 (*left*). The system conceptualizes those categories using the extracted object views. Afterwards, the system is tested by the remaining 12 scenes captured from different viewpoints (i.e., shown by red cameras). The system could recognize all objects properly by using the knowledge learned from the first scene. Some misclassifications also occurred throughout the demonstration. The underlying reason was that, at some points, the object tracking could not track the object properly and the distinctive parts of the object were not included in the object's point cloud. This evaluation illustrates the process of learning object categories in an open-ended fashion. A video of this demonstration is available online at: https://youtu.be/zjucGaAwnTE

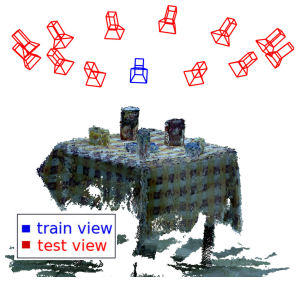


Fig. 5. Experimental setup using Imperial College dataset [17].

VI. CONCLUSION

In this paper, we have tackled the problem of open-ended object category learning and recognition by proposing a new object representation, which is suitable for both fine-grained and basic-level object categorization. In particular, we described each object based on a set of *general latent visual topics* and a set of *category-specific* dictionaries. An extensive set of experiments was carried out to assess the performance of the proposed approach. Experimental results show that the overall classification performance obtained with the proposed approach is clearly better than the best accomplishments achieved with the state-of-the-art methods. Moreover, the best scalability was obtained with the proposed approach, followed by the Local-LDA [5]. Concerning computational time, the best result was obtained with BoW [18], immediately followed by the Open-Ended LDA approach [19]. A real demonstration proved that the agent could learn new categories from very few examples in an incremental and open-ended manner.

REFERENCES

[1] M. Oliveira, L. Seabra Lopes, and et al., "3D object perception and perceptual learning in the RACE project," *Robotics and Autonomous Systems*, vol. 75, Part B, pp. 614 – 626, 2016.

[2] J. Young, L. Kunze, V. Basile, E. Cabrio, N. Hawes, and B. Caputo, "Semantic web-mining and deep vision for lifelong object discovery," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2774–2779.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proc. of IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy*, 2017.

[5] S. H. Kasaei, A. M. Tome, and L. Seabra Lopes, "Hierarchical object representation for open-ended object category learning and recognition," in *Advances in Neural Information Processing Systems (NIPS 2016) 29*, 2016, pp. 1948–1956.

[6] S. H. Kasaei, M. Oliveira, G. H. Lim, L. S. Lopes, and A. M. Tomé, "Interactive open-ended learning for 3D object recognition: An approach and experiments," *Journal of Intelligent & Robotic Systems*, vol. 80, no. 3-4, pp. 537–553, 2015.

[7] M. Ullrich, H. Ali, M. Durner, Z.-C. Márton, and R. Triebel, "Selecting cnn features for online learning of 3D objects," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 5086–5091.

[8] S. H. Kasaei, J. Sock, L. S. Lopes, A. M. Tomé, and T.-K. Kim, "Perceiving, learning, and recognizing 3D objects: An approach to cognitive service robots," in *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18)*, 2018.

[9] M. Oliveira, L. S. Lopes, G. H. Lim, S. H. Kasaei, A. D. Sappa, and A. M. Tomé, "Concurrent learning of visual codebooks and object categories in open-ended domains," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 2488–2495.

[10] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 623–634, 2014.

[11] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1713–1725, 2016.

[12] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with latent dirichlet allocation," in *International conference on artificial intelligence and statistics*, 2009, pp. 65–72.

[13] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 5, pp. 433–449, May 1999.

[14] S. Chakraborty and N. Nagwani, "Analysis and study of incremental k-means clustering algorithm," in *High performance architecture and grid computing*. Springer, 2011, pp. 338–341.

[15] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, 2007.

[16] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 1817–1824.

[17] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6D object pose and predicting next-best-view in the crowd," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3583–3592.

[18] S. H. Kasaei, M. Oliveira, G. H. Lim, L. S. Lopes, and A. M. Tomé, "Towards lifelong assistive robotics: A tight coupling between object perception and manipulation," *Neurocomputing*, vol. 291, pp. 151–166, 2018.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[20] S. H. Kasaei, L. Seabra Lopes, and A. M. 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)*. IEEE, 2018, pp. 6806–6812.