Incremental Object Model Learning from Multimodal Human-Robot Interactions

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Learning object models in the wild from natural human-robot interactions is essential for robots to operate in real environments. Natural human interactions are in essence multimodal, including among others language and gestures. The main contribution of this paper is the development and evaluation of an incremental learning algorithm that uses data from such interactions. Our experiments show the first results within this area and confirm the challenges of the task.

1 Introduction

Models trained offline on large datasets cannot, in general, address some challenges of real data in home environments. One example is the long-tail distribution, i.e., objects that appear rarely and for which few or none training samples exist in common databases. Another example is the changing nature of the environments, with new objects appearing, e.g. food products that did not exist when the large training datasets were created. In order to address these and other cases, robotic learning should be incremental. Moreover, a key aspect in service robotics is a comfortable and intuitive human-robot interaction. Such interaction is needed to capture data to update the world models incrementally, from the user's knowledge and behavior, and in a natural manner. We believe the best interaction is natural language and gestures, similarly to how the user would teach something to another person.

This paper addresses incremental learning of object models from natural human-robot interaction. The human should be able to teach unknown objects to the robot, so that the robot can identify them later on. Our approach (Figure 1) is based on [2] and brings specific contributions at the 3 main steps:

Figure 1: Overview of our approach. A human user teaches a robot new objects through natural interactions (e.g., pointing to it). The robot recognizes the type of interaction, finds the corresponding object region on its camera views and updates the object model incrementally with that data.

Multimodal Interaction Recognition. An accurate identification of the human-robot interactions is a key aspect, as the strategy to find object patches, needed for training, depends on it. Differently from [2], we incorporate user skeleton detection [4] to guide the hand search.

Target Object Detection. For each interaction type we select the image patches that are likely to contain the target object. We use a combination of the segmentation given by MaskRCNN [9] and the superpixel segmentation proposed in [2].

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

Incremental Learning. This is our main contribution. Incremental learning was not addressed in [2], since they focused mostly on the previous two aspects. The candidate patches obtained in the previous step are used as training samples. We propose an approach based on incremental clustering and K-Nearest Neighbour for classification.

2 Related work

Very related to our work, Pascuale et al. [18] uses CNN features and SVM for visual recognition. The training data consists of egocentric images where a human presents an object in front of the robot. Camoriano et al. [3] uses such data and presents a variation of Regularized Least Squares for incremental object recognition. In mobile robotics, we find multiple examples that propose how to incrementally adapt environment visual models as the robot moves. These approaches are often based on Gaussian mixture models that can be easily updated and maintained to recognize regions of interest for the robot [6, 19]. Yao et al. [24] proposed an incremental learning method, that continually updates an object detector and detection threshold, as the user interactively corrects annotations proposed by the system. Kuznetsova et al. [15] investigated incremental learning for object recognition in videos. Vatakis et al. [22] shows multimodal recording approach similar to ours, but their dataset's goal was to capture user reactions to stimuli with objects or images in a screen.

In recent years, significant advances have been made in the field of incremental learning. Works like Aksoy et al. [1] are able to incrementally learn semantic event chains (SECs) extracted from actions using human demonstration. The most classic works presented variations or combinations with k-means clustering algorithm. Murty et al. [17] combines k-means with multilevel representation of the clusters. Likas et al. [16] presents a global algorithm that adds a new cluster and dynamically updates the others by applying the k-means algorithm multiple times. Other approaches apply a data transformation based on self-organizing maps (SOM) Neural Networks. [7] presents an online unsupervised system with an incremental update of a Neural Network based on SOM (SOINN). [23] presents a variant of the Self-Organizing Incremental Neural Networks that incrementally transform the nodes in the layers of the SOINN using the local distribution. [8] uses SOM to reduce the dimensionality of the data, but it needs to keep all the data in memory for re-training. [10] presents a work that combines the SOINN data transformation with SVM for classification.

In robotics, we find situations where the robot interacts directly with the scene, e.g., moving an object, to build an incremental object model [13, 5, 12, 14, 20]. Our approach is complementary to these works, as human interaction is needed in real scenarios, e.g., if the object is out of robot's reach.

3 Incremental Object Model Learning from Interactions

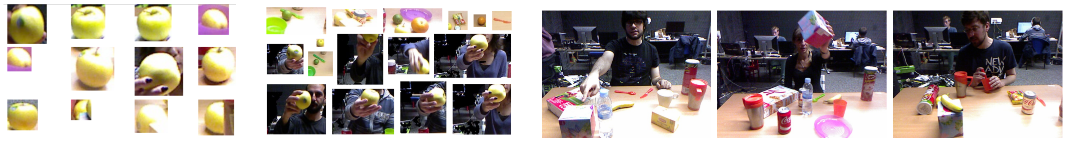
Our approach enables a robot to learn object models incrementally, while limiting the size of the stored data. The proposed approach selects and stores representative *object views* (image patches) for each object, selected from the input candidate patches obtained following the strategy from [2].

3.1 Object model and descriptors.

Our database consists of a set of descriptors for each representative object view. Each of these descriptors is the centroid of a database cluster, and an *object model* will be composed of several of these clusters. We consider descriptors that are reasonably small and fast to compute, since our system is designed for robotic platforms, where computation is typically limited. Besides, for an illustration of typical common object patches in robotic settings, Figure 2 shows a few examples from MHRI dataset [2]. Those examples show the typical low resolution and high clutter, even in manually cropped patches. Our goal is to recognize common objects in this type of realistic views, for which we evaluate several descriptors (detailed in the experiments): common hand-crafted features and deep learning based features (i.e., final layer outputs from several well known classification CNNs).

3.2 Incremental Object Learning

The processing of new incoming object views, either to update the object models or to perform recognition, is as follows.



(a) (b) (c)

Figure 2: Examples of MHRI data [2]. (a) *Manually Cropped* and (b) *Automatically segmented* patches from a sample object (apple). (c) Interaction types (from left to right: *Point*, *Show* and *Speak*).

Model initialization and incremental update. Given a new object view v and its label l, we compute its descriptor d^v and create a new cluster $C[\hat{v}]$ with d^v as centroid and l as associated label.

If label l **does not exist** in the database, l is added to the database initializing a new object.

If label l **already exists** in the database, existing clusters evolve and update their centroids (representative descriptors), following incremental clustering ideas. The total number of classes is not limited but, in order to avoid unlimited growing, the subset for clusters within each class is limited by a predefined size. If l has reached this maximum number of clusters, we run an alternating strategy that is repeated every $n+1$ updates to a certain label l: 1) *For the first* n *updates* to label l, our algorithm computes the distances among all clusters associated with label l (C_l), in order to find the closest and the furthest pairs among them. To compare two clusters we use the distance between their centroids. The closest *pair* of clusters are merged, updating the centroid and increasing its positive count with one. Oppositely, the furthest pair of clusters, receive a negative vote. 2) *For the* $n+1$ *update* to label l, the cluster with the worst score (i.e., more negative votes) is replaced by the new singleton cluster.

Additionally, we tested random and minimum distance as another criterion for this cluster reorganization step. But the proposed method gives a better performance (accuracy of 13% against 11% and 10% respectively), and prevents too much similarity or disparity among clusters from the same label.

Recognition. To classify a new object view v into the existing classes, we simply follow a k-Nearest Neighbor (k-NN) approach (in our tests, $k = 3$). The distance between current view descriptor d^v and each existing model cluster is computed, and the view is assigned the label according to the most frequent label from the closest k neighbours found.

4 Experimental Validation

All our experiments use the Multi-modal Human-Robot Interaction (MHRI) dataset [2]. This dataset captures the most common natural interactions for teaching object classes to a robot: *Point*, *Show* and *Speak*. It contains clips from 10 users doing 3 types of interaction with 10 objects from a pool of 22 objects. Our focus is on exploring incremental learning strategies for the object model part of the pipeline proposed together with the dataset. However, during our implementation (built on the code provided by authors in [2]) we have also improved their interaction recognition and their target object detection modules, as described in the introduction. This improved implementation will be released to the community.

Incremental Learning Module Evaluation: This experiment evaluates the proposed incremental learning strategy decoupled from the quality of the data, i.e., we use **manually segmented patches** from MHRI dataset (670 patches from 22 classes, approx. 30 patches per class and 67 patches per user). Figure 2(a) shows examples of such patches. We do 10-fold cross validation, each fold keeping one user for testing and the rest of users for training. The supplementary material includes detailed results with additional baselines and variations. Table 1(a) only shows the most insightful results.

Model size limit. We considered different cluster size limits (including no-limit). After a cluster-size limit of 20, we observed that the accuracy did not improve substantially, and hence it is reasonable to implement such limit in constrained platforms.

Different patch descriptors. We show the best result for hand-crafted features (color histograms HC_{RGB}) and for deep learning based features ($DenseNet_4$, output of the Dense Block 4 of pre-trained DenseNet [11]). HC_{RGB} provided the highest accuracy, surprisingly at first sight, but it can be explained by looking at the MHRI data: objects with distinctive colors and poor texture. All

Table 1: Average accuracy for object recognition using different approaches with MHRI data.

Patches:	(a) *Manually Cropped* (clean)	(b) *Automatically Segmented* (noisy)
Incremental k-NN+HC_{RGB} (Ours)	28.0* / 31.4**	9.0* / 13.2**
Incremental k-NN+$DenseNet_4$ (Ours)	18.28* / 21.17**	5.5* / 5.6**
Offline k-NN+HC_{RGB}	30.2	13.4
Offline SVM+HC_{RGB} [2]	34.8	7.95
Offline CNN (*Inception*-finetuned)	59.3	17.5

* 50% of data processed by the incremental system. ** 100% of data processed by the incremental system

evaluated descriptors, except HC_{RGB}, fire around high-gradient regions. And the CNNs considered, are pre-trained with very different type of images (ImageNet) with wider FOV images, hence most learned features probably do not apply in our patches. This just confirms the issues with domain change using CNN-based strategies with this dataset already discussed in detail in [2].

Related offline baselines. The best performing offline baseline is an Inception V3 model [21], pre-trained on ImageNet and fine-tuned with the training set of the *Manually-cropped* patches. This is an upper bound for the performance worth showing as reference. However, it is not suitable for incremental learning, since the update data we get from a few user interactions is not enough to fine-tune further the net. The most significant observation is that our proposed *Incremental k-NN* strategy gets similar performance to an *offline k-NN* that uses all the data at once. This validates the incremental approach and verifies the strategy to limit the cluster size is not harming the performance.

Validation of the full pipeline: This experiment uses object **patches extracted automatically** from interactions for training and testing. Figure 2(b) shows examples of these *automatic patches*, with significantly worse quality that *manual patches*. This increases the challenge but brings the experiment closer to a system running in the wild. The supplementary material includes more results with additional baselines and variations. Table 1(b) shows the most insightful results, discussed next.

Incremental k-NN. The incremental system we propose is evaluated with a 10-fold cross-validation, where each fold corresponds to a user, and set to the best performing configuration from previous experiment (HC_{RGB} descriptor and model size limit 20). Besides the challenge from using automatically segmented patches, note that each user manipulates a different subset of the object pool, i.e., at some points for some of the folds (depending on which user data has been fed to the incremental system), there were no training examples for some of the test data objects. Since users do not have clips with all the objects in the pool, *Incremental k-NN* needs to process several users (4 in our experiments) to reach a reasonable performance. The average accuracy of our incremental k-NN approach is again similar to an offline k-NN, but storing a significantly lower amount of data.

Comparison with offline baselines. Up to our knowledge there is not another available end-to-end system of similar characteristics to ours. Therefore, we show as reference the results of the same offline approaches as in previous experiment. We can see all approaches suffer a significant decrease in performance with respect to what they reached training with *Manual patches* in previous experiment. This is not surprising and confirms the challenging set up we are working with. Our incremental approach also suffers a decrease in performance but it is able to outperform the baseline of [2] using only 50% of the data. Note that in this case the other offline baselines are not much better than our incremental approach, which highlights the challenging data and setup considered and leaves open research problems in learning for service robotics.

5 Conclusions

This paper presents the first complete approach for incremental object learning using multimodal data from natural Human-Robot interaction. The pipeline is based on [2], improving all its stages, proposing an incremental learning approach and presenting results on a public database. Our novelty is on the integration of several modules that facilitate the use of natural language and gestures for incremental robot learning. Our main insights are 1) the domain change is critical in this scenario, and 2) although we reach a reasonable performance there are still considerable challenges, justifying the relevance of the topic for future research. We believe that the most relevant one is the exploration of more sophisticated incremental learning methods, particularly those that are robust to noisy data.

References

[1] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter. Model-free incremental learning of the semantics of manipulation actions. *Robotics and Autonomous Systems*, 71:118 – 133, 2015. Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence.

[2] P. Azagra, F. Golemo, Y. Mollard, M. Lopes, J. Civera, and A. C. Murillo. A multimodal dataset for object model learning from natural human-robot interaction. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6134–6141, Sept 2017 `http://robots.unizar.es/IGLUdataset/`.

[3] R. Camoriano, G. Pasquale, C. Ciliberto, L. Natale, L. Rosasco, and G. Metta. Incremental object recognition in robotics with extension to new classes in constant time. *arXiv preprint arXiv:1605.05045*, 2016.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *IEEE Int. Conf. on Robotics and Automation*, pages 48–55, May 2009.

[6] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. Bradski. Self-supervised monocular road detection in desert terrain. In *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.

[7] S. Furao and O. Hasegawa. An incremental network for on-line unsupervised classification and topology learning. *Neural Networks*, 19(1):90 – 106, 2006.

[8] A. Gepperth and C. Karaoguz. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5):924–934, Oct 2016.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[10] A. Hebboul, F. Hachouf, and A. Boulemnadjel. A new incremental neural network for simultaneous clustering and classification. *Neurocomputing*, 169:89 – 99, 2015. Learning for Visual Semantic Understanding in Big Data ESANN 2014 Industrial Data Processing and Analysis.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[12] P. Iravani, P. Hall, D. Beale, C. Charron, and Y. Hicks. Visual object classification by robots, using on-line, self-supervised learning. In *IEEE Int. Conf. on Computer Vision Workshops*, pages 1092–1099, 2011.

[13] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *IEEE Int. Conf. on Robotics and Automation*, pages 1377–1382, 2009.

[14] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3D object models using next best view manipulation planning. In *IEEE Int. Conf. on Robotics and Automation*, pages 5031–5037, 2011.

[15] A. Kuznetsova, S. J. Hwang, B. Rosenhahn, and L. Sigal. Expanding object detector's horizon: incremental learning framework for object detection in videos. In *Computer Vision and Pattern Recognition*, pages 28–36. IEEE, 2015.

[16] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[17] M. N. Murty and G. Krishna. A hybrid clustering procedure for concentric and chain-like clusters. *International Journal of Computer & Information Sciences*, 10(6):397–412, 1981.

[18] G. Pasquale, C. Ciliberto, F. Odone, L. Rosasco, L. Natale, and I. dei Sistemi. Teaching iCub to recognize objects using deep convolutional neural networks. *Proc. Work. Mach. Learning Interactive Syst*, pages 21–25, 2015.

[19] J. Rituerto, A. C. Murillo, and J. Kosecka. Label propagation in videos indoors with an incremental non-parametric model update. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2383–2389, Sept 2011.

[20] J. Sinapov, C. Schenck, and A. Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE Int. Conf. on Robotics and Automation*, pages 5691–5698, 2014.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[22] A. Vatakis and K. Pastra. A multimodal dataset of spontaneous speech and movement production on object affordances. *Scientific Data*, Jan 2016.

[23] Y. Xing, X. Shi, F. Shen, K. Zhou, and J. Zhao. A self-organizing incremental neural network based on local distribution learning. *Neural Networks*, 84:143 – 160, 2016.

[24] A. Yao, J. Gall, C. Leistner, and L. Van Gool. Interactive object detection. In *Computer Vision and Pattern Recognition*, pages 3242–3249. IEEE, 2012.