Making Sense of Indoor Spaces Using Semantic Web Mining and Situated Robot Perception

Making Sense of Indoor Spaces Using Semantic Web Mining and Situated Robot Perception

Jay Young, Valerio Basile, Markus Suchi, Lars Kunze, Nick Hawes, Markus Vincze, Barbara Caputo

Making Sense of Indoor Spaces Using Semantic Web Mining and Situated Robot Perception

Jay Young[1], Valerio Basile[2], Markus Suchi[3], Lars Kunze[4], Nick Hawes[1], Markus Vincze[3], and Barbara Caputo[5]

[1] The University of Birmingham, United Kingdom
[2] Université Côte d'Azur, Inria, CNRS, I3S, France
[3] Technische Universität Wien, Austria
[4] Oxford Robotics Institute, Dept. of Engineering Science, University of Oxford, UK
[5] Università di Roma - Sapienza, Italy

Abstract. Intelligent Autonomous Robots deployed in human environments must have understanding of the wide range of possible semantic identities associated with the spaces they inhabit – kitchens, living rooms, bathrooms, offices, garages, etc. We believe robots should learn this information through their own exploration and situated perception in order to uncover and exploit structure in their environments – structure that may not be apparent to human engineers, or that may emerge over time during a deployment. In this work, we combine semantic web-mining and situated robot perception to develop a system capable of assigning semantic categories to regions of space. This is accomplished by looking at web-mined relationships between room categories and objects identified by a Convolutional Neural Network trained on 1000 categories. Evaluated on real-world data, we show that our system exhibits several conceptual and technical advantages over similar systems, and uncovers semantic structure in the environment overlooked by ground-truth annotators.

1 Introduction

Many tasks in Human-Robot Interaction (HRI) scenarios require autonomous mobile service robots to relate to objects and places (or rooms) in their environment at a semantic level. This capability is essential for interpreting task instructions such as "Get me a mug from the kitchen" and for generating referring expressions in real-world scenes such as "I found a red and a blue mug in the kitchen, which one should I get?" However, in dynamic, open-world environments such as human environments, it is simply impossible to pre-program robots with the required knowledge about task-related objects and places in advance. Hence, they need to be equipped with learning capabilities that allow them to acquire knowledge of previously unknown objects and places online. In previous work, we demonstrated how knowledge about perceived objects can be acquired by mining textual resources [8] and image databases on the Semantic Web [9]. In this work, we focus on *knowledge about places* and investigate ways

of acquiring it using web mining and situated robot perception. In particular, we aim to learn the semantic categories of places observed by an autonomous mobile robot in real-world office environments.

When mobile service robots are deployed in human-inhabited locations such as offices, homes, industrial workplaces and similar locations, we wish them to be equipped with ways of learning and the ability to extend their own knowledge on-line using information about the environment they gather through *situated experiences*. This too is a difficult task, and is much more than just a matter of data collection. Some form of *semantic* information is desirable too. We expect that structured and semi-structured Web knowledge sources such as DBPedia and WordNet [2] to answer some of these questions. By linking robot knowledge to entries in semantic ontologies, we can begin to exploit rich knowledge-bases to facilitate better robot understanding of the world.

One data source of interest to us is ImageNet, which is a large database of categorised images organised using the WordNet lexical ontology. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [3] has in recent years produced machine learning tools trained on ImageNet for object detection and image classification. Of particular interest to us are *deep learning* based approaches using Convolutional Neural Networks, trained on potentially thousands of object categories [4]. This approach raises the question of how well such predictors perform when queried with the challenging image data endemic to mobile robot platforms, as opposed to the cleaner, and higher-resolution, data they are typically trained and evaluated on. This domain adaptation problem is a major difficulty in using these state-of-the-art vision techniques on robots. Using vision techniques with (ever-growing) training sets the size of ImageNet, will allow us to extend a robot's knowledge base far beyond what it can be manually equipped with in advance of a deployment.

In this paper we document our work using the technologies mentioned so far towards enabling a mobile robot to learn the semantic categories associated with different regions of space in its environment. To do this, we employ large-scale object recognition systems to generate semantic label hypotheses for objects detected by robots in real-world environments. These hypotheses are linked to *structured, semantic knowledge bases* such as DBPedia and WordNet, allowing us to link a robot's situated experiences with higher-level knowledge. We then use these object hypotheses to perform text-mining of the semantic web to produce further hypotheses over the semantic category of particular regions of space.

To summarise, this paper makes the following contributions:

- an unsupervised approach for learning semantic categories of indoor spaces using deep vision and semantic web mining;
- an evaluation of our approach on real-world robot perception data; and
- a proof-of-concept demonstration of how knowledge about semantic categories can be transferred to novel environments.

2 Previous Work

Space categorisation for mobile robots is an extensive, well-studied topic, and one which it would be impossible to provide an in-depth review of in the space available. For this, we would reccomend the work of [13], which provides a thorough survey of the wider field of robot semantic mapping to-date. The majority of work in the area of space categorisation utilises *semantic cues* to identify and label regions of space such as offices, hallways, kitchens, bathrooms, laboratories, and the partitions between them. One of the most commonly used semantic cues is the presence of objects, and as this is also the semantic cue we use, we will focus on this area of the work.

The work of [10] realises a Bayesian approach to room categorisation, and builds a hierarchical representation of space. This hierarchy is encoded by the authors, who admit that their own views and experiences in regards to the composition of these concepts could bias the system. In further work, the same authors [11] provide a more object-focused approach to space classification, however this again required the development and evaluation of a knowledge base linking objects to room types. The work of Pronobis and Jensfelt [12] is significant in this area in that it integrates heterogeneous semantic cues, such as the shape, size and appearance of rooms, with object observations. However, their system was only capable of recognising 6 object types and 11 room categories, which again required the gathering and annotation of much training data, and it is unclear how well this generalises to new environments and how much re-training would be required. Similar systems [14] exhibit the same pitfalls. The work of Hanheide [15] on the Dora platform realises a robot system capable of exploiting knowledge about the co-occurence of objects and rooms. This is facilitated by linkage to the *Open Mind Indoor Common Sense* database, and is used for space categorisation and to speed up object search by exploiting semantic relations between objects and rooms.

We argue that our approach exhibits several technical and conceptual advantages over other pieces of work in this area:

- The categorisation module requires no robot perceptual data collection or training, and works fully on-line.
- The system is domain agnostic, not fitted to particular types of environments, room structures or organisations.
- We use existing, mature, tried-and-tested semantic ontologies, and as such there is no knowledge-engineering required by the system designer to use this information.
- The use of large-scale object recognition tools mean we are not limited to a small number of objects, and the use of text-mining means we are not limited to a small number of room categories.
- The relations between objects and room categories are derived *statistically* from text mining, rather than being encoded by the developer or given by an ontology.

These key points lead to a novel way of solving the problem of space classification on mobile robots.

3 Approach Overview

Fig. 1: Overview.

We use a robot platform to observe the environment at various waypoints specified in its environment. The robot is provided with a SLAM map of its environment, and a set of waypoints within this map. At each of these places the robot perceives its surroundings by taking multiple views at different angles (360°). The different views of the robot are aligned and integrated into a consistent environment model in which object candidates are identified and clustered into groups according to their proximity. For each object candidate, we predict its class by using its visual appearance as an input to classifiers trained on a large-scale object database, namely ImageNet. Based on the set of labelled (or classified) object candidates which are in the same group, we perform a web-based text-mining step to classify the region of space constrained by a bounding polygon of the group of objects.

In the following, we describe the individual components in more detail.

4 Object Category Recognition

Our aim is to identify the semantic labels most strongly associated with a particular point in a robot's environment by looking at the kinds of objects that are visible from that point. As such, it is crucial for a robot to be able to recognise the objects that inhabit its environment. It is typical in robotics that object recognition is facilitated by a training step prior to deployment [12,15] (though unsupervised approaches do exist [1]) whereby selected objects from the robot's environment are learned and later re-recognised and used for space categorisation. The advantage of this is that the robot learns to recognise objects using models trained using its own sensors and situated conditions, however it also means that we must anticipate which objects a robot is likely to encounter so as

to determine which ones to learn and which to ignore. This process can also be very time-consuming and error-prone.

Previous work [9] has used Convolutional Neural Networks (CNNs) trained on large image databases such as ImageNet, which provide databases of several million images, for object recognition on a mobile robot. Results can vary, and this is because the images used to train ImageNet-sourced CNNs possess very different characteristics to those images observed by robots – robot data is often noisy, grainy and typically low-resolution, and is exasperated by the difficulties robots have in getting close to objects, especially small ones. One cause of this is what is known as the *domain adaptation problem*, where the features learning mechanisms discover from their high-resolution training data do not robustly and reliably map on to lower-resolution, noise-prone spaces. This is an active, ongoing area of research in the computer vision community, the solution to which holds the key to generic, off-the-shelf object recognition for mobile robots.

We evaluated a set of state-of-the-art CNNs trained on ImageNet on a sample (1000 object images) from one of our robot datasets. We measure our accuracy using a WUP similarity score [5], which calculates the semantic relatedness of the ground-truth concept types against the concept predicted by the CNN by considering their depth of their lowest common super-concept in the WordNet ontology. A WUP score of 1.0 means two concepts are identical. The concepts *Dog* and *cat*, for instance, have a WUP relatedness score of 0.86. To compare, we also built a wrapper for the Google Web Vision API, that mapped its output to the WordNet ontology. We evaluated against Google Web Vision, the GoogleNet CNN, and the AlexNet and ResNet152 CNNs. Our results were 0.392, 0.594, 0.590 and 0.681 respectively, given as average WUP score over the randomly sampled 1000 images from our labelled robot dataset. As such, we chose the ResNet152 model to work with [17].

4.1 Scene Segmentation

In order to identify objects we must first have an idea about where they are in the environment. To generate object location hypotheses we make use of our own implementation of the RGB-D depth segmentation algorithm of [16]. This is a patch-based approach, which clusters locally co-planar surfaces in RGB-D point clouds. These initial surfaces are geometrically modeled into planes and non-uniform rational B-splines using a best fit approach. The adjacency relation between those models yield a graph and by applying a graph-cut algorithm refine the segmentation. Given an observation of a scene from the robot, this algorithm returns a set of segmented candidate objects from the scene. From there, we perform basic filtering for instance to filter out objects that are too small or too dark, and are likely to be erroneously segmented environmental noise. We can then extract the 2D bounding-box around the objects to be passed directly to the object recognition system.

5 Text Mining

There has been recent work towards developing a Semantic Web-Mining component for mobile robot systems [8,9] which we make use of. This component provides access to object- and scene-relevant knowledge extracted from Web sources, and is accessed using JSON-based HTTP requests. The structure of a request to the system describes the objects that were observed in a scene, and has been used to identify unknown objects given their context. In this case the service computes the *semantic relatedness* between each object included in the co-occurrence structure and every object in a large set of candidate objects (the *universe*) from which possible concepts are drawn from. This semantic relatedness is computed by leveraging the vectorial representation of the DBpedia concepts provided by the NASARI resource [6]. The NASARI resource represents BabelNet concepts as a vector in a high-dimensional geometric space. In this case using Wikipedia as source corpus. The system computes the aggregate of the relatedness of a candidate unknown object to each of the scene objects contained in the query, returning a ranked list of object label candidates based on relatedness. We re-work this same approach to instead return ranked relatedness distributions over *room categories* given a *set of observed objects*. We used the following room categories: Kitchen, Office, Eating Area, Garage, Bathroom. The system then provides a distribution over these categories for input sets of objects.

6 Experiments and Results

We employ two datasets of observations taken by our robot during two long-term (3 months) deployments in two separate office environments a year apart. The first dataset was labelled by a human to produce 3800 views of various objects, with the data collection methodology following the approach of Ambruş et. al [7]. The robot is provided with a map, and a set of waypoints in the map that it visits several times per day, performing full 360° RGB-D scans of the environment at those points. The second dataset is as-yet unlabelled.

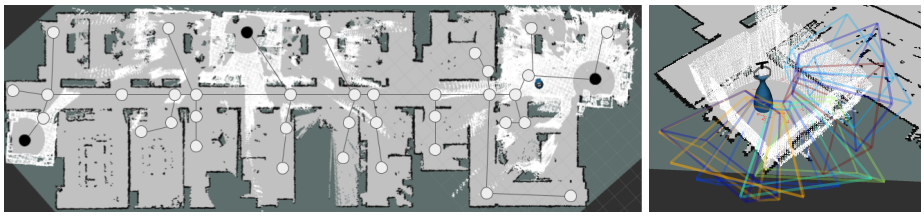


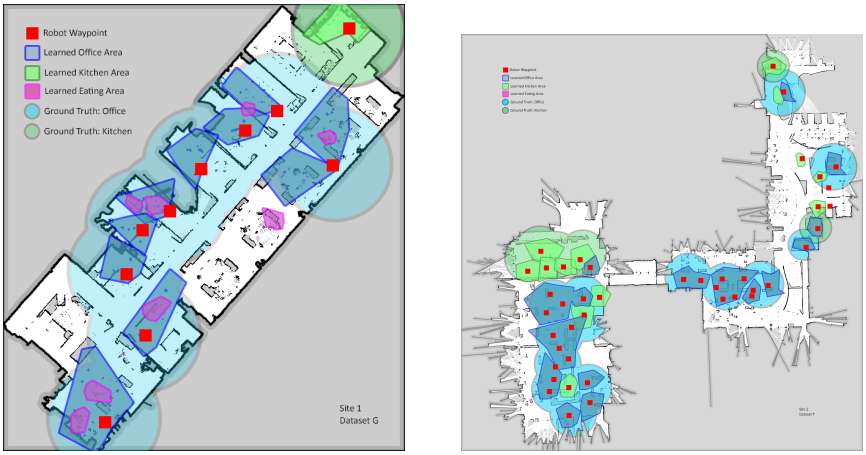
Fig. 2: Experimental Setup at G. **Left:** A robot makes 360° scans at several predefined waypoints in its environment. **Right:** robot plans views to investigate parts of the mapped environment.

We perform two main experiments – first, we demonstrate the results of our approach on the first, human-labelled dataset gathered from site 1 (dataset G). Since this is hand-labelled it gives us access to a representation of the objects encountered by the robot under ideal conditions – assuming no segmentation errors, and perfect object recognition. First, we sample the objects observed at each waypoint over the period of the deployment by selecting the top-n occurring objects, here using $n == 30$. From here we perform Euclidean Clustering to group objects together, producing clusters of those objects that appear within $0.5m$ of one-another.

Each of these clusters is then incrementally sent to our text-mining module. In return, we receive a distribution over room categories at those points in space. After all clusters have been processed we perform a round of merging, coalescing any clusters that possess centroids within a $1.5m$ of one-another, and which share the same top-ranked category. From here, we can use these new clusters to calculate bounding polygons to produce larger, categorised spatial regions. For a more intuitive representation, we found it helpful to include an inflation parameter for this – because we would like to categorise the area *around* an object or set of objects, which we expect is better served by a geometrical bounding area around objects rather than treating them as points. We apply a bounding area of $1.5m$ around objects.

In our second experiment, we perform the exact procedure as described above on data gathered from site 2 (dataset T), however the input to the system takes the form of dynamically segmented objects using the segmentation procedure described previously, and using object hypotheses from the ImageNet-based CNN approach. Since this dataset is significantly larger, we sampled from it an equal number of observations per waypoint (4), providing us with roughly 2800 individual RGB-D clouds of scenes of the environment. Segmenting these resulted in $85,000$ segments, however we applied a standard filtering by ignoring any segments that were more than $2m$ away from the robot base, which filtered the set of segments down to roughly $24,000$.

To evaluate our results, we provided each of the clusters of objects to five human annotators, and asked them to identify the room categories they believed to be most closely related to the set of objects. This was done without visual information on the appearance of the objects or the environment in which they were found, in the first experiment at site 1 we achieved an agreement between the annotators and the system of 74%. In the second experiment at site 2, we achieved an agreement of 80% between annotators and our system. In a second round of evaluation, a different set of seven annotators were provided images observed by the robot at each waypoint, and asked to identify the likely room categories displayed in the images from the same set of candidate rooms provided to the robot. We apply these ground-truth labels to the areas of space around each waypoint. This allows us to compare these ground-truth category labels with the labels suggested by our system. The results are shown in Figure 3. On the map, dark blue polygons represent regions learned by our system, red squares indicate the waypoints where the robot took observations, and light coloured



(a) Site 1, G dataset (b) Site 2, T dataset

Fig. 3: Space categorisation results from both sites, showing learned and categorised regions and ground-truth annotations. A zoomed view is recommended.

circles indicate the ground-truth label of the space around each waypoint – human annotators agreed on labels for these areas, so there is no variance.

7 Discussion

In our results from site 1, the system categorised three region types – kitchen, office and eating area. Our ground-truth labellers, given the same list of candidate rooms as the robot, only labelled kitchen and office areas. All of the office and kitchen areas learned by the system fall into the corresponding areas labelled by the human annotators, and represent a sub-section of that space. These were labelled by detecting objects such as filing cabinets, computer equipment, printers, telephones and whiteboards, which all ultimately most strongly correlated with the office room category. But where do the eating areas come from? These areas were labelled by detecting objects such as water bottles, coffee cups and mugs on the desks and cabinets of workers in the deployment environment. These objects were typically *surrounded* by office equipment. While comparing these region labels to our ground-truth data would suggest the answer is wrong, we believe that this captures a more finely-grained semantic structure in the environment that does in fact make sense. While the regions themselves may not, to a human, meet the requirements for a dining area, the objects encompassed within them are far more closely linked in the data with eating areas and kitchens than they are with computer equipment and stationary, and so the system annotates these regions differently.

At site 2 we see that the robot did not learn these characteristic eating area regions. While inspection of the data shows that many desks do exhibit the same

structure of having mugs, cups and bottles on them in certain areas, the object recognition system used in the second set of experiments failed to correctly identify them. These objects are typically small, and difficult for a mobile robot to get close to. The results for the second dataset are also more noisy – there are misclassified regions. These were caused primarily by object recognition errors, themselves compounded by segmentation errors and sensor noise. To filter these out, we included a filter on system that ignored any classification result that came back with a confidence below 0.1 – ignoring those objects completely filtered out around $18,000$ segments.

Our system is ultimately limited by its reliance on objects to generate hypotheses for space classification. This means that our approach is unable to categorise areas of space such as corridoors or hallways. However it is intended to work as a component of object-search systems, so perhaps this is not necessary at this stage. To illustrate this, we built a query interface for the system which takes an arbitrary object label and suggests an area of space where the object can be found, ranking results using the semantic relations of the object with the categories learned at each region. This allows a robot to generate priors over possible locations of objects it has never seen before, and we view as the first step towards *unknown object search*.

There are many different possible representations for the data our system generates. We opted for a clustering and bounding-polygon based approach in order to most clearly visualise our results, but other approaches could be used such as flood-fill algorithms, heat-maps or potential fields. Choice of representation should be informed by the task that is intended to make use of the information.

8 Conclusion

In this work we presented a robot system capable of categorising regions of space in real-world, noisy human-inhabited environments. The system used concepts in a lexical ontology to represent object labels, and harnessed this representation to mine relations between observed objects and room categories from corpora of text. Transferring these relations back to the real-world, we used them to annotate the robot's world with polygons indicating specific semantic categories. We found that the system was largely able to discover and categorise regions similar in area to human annotators, but was also able to discover some structure overlooked by those annotators.

Acknowledgments

The research leading to these results has received funding from EU FP7 grant agreement No. 600623, STRANDS, and CHIST-ERA Project ALOOF.

References

1. T. Faeulhammer, et. al.: "Autonomous learning of object models on a mobile robot," *IEEE RAL*, vol. PP, no. 99, pp. 1–1, 2016.

2. A. Kilgarriff and C. Fellbaum, "Wordnet: An electronic lexical database," 2000.
3. O. Russakovsky, et. al. "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
4. A. Krizhevsky, et. al. "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
5. Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *ACL*, ser. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 133–138.
6. J. Camacho-Collados, et. al. "Nasari: a novel approach to a semantically-aware representation of items." in *HLT-NAACL*, R. Mihalcea, J. Y. Chai, and A. Sarkar, Eds. The Association for Computational Linguistics, 2015, pp. 567–577.
7. R. Ambruş,et. al , "Meta-rooms: Building and maintaining long term spatial models in a dynamic world," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1854–1861.
8. J. Young, et. al , "Towards lifelong object learning by integrating situated robot perception and semantic web mining," in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2016.
9. J. Young, et. al , "Semantic Web-Mining and Deep Vision for Lifelong Object Discovery" in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
10. Vasudevan, Shrihari, and Roland Siegwart. "Bayesian space conceptualization and place classification for semantic maps in mobile robotics." *Robotics and Autonomous Systems* 56.6 (2008): 522-537.
11. Vasudevan, Shrihari, et. al *"Cognitive maps for mobile robotsan object based approach."* Robotics and Autonomous Systems 55, no. 5 (2007): 359-371.
12. Pronobis, Andrzej, and Patric Jensfelt. "Large-scale semantic mapping and reasoning with heterogeneous modalities." *IEEE International Conference on Robotics and Automation*, 2012.
13. Kostavelis, Ioannis, and Antonios Gasteratos. "Semantic mapping for mobile robotics tasks: A survey." *Robotics and Autonomous Systems* 66 (2015): 86-103.
14. Zender, Hendrik, et. al *"Conceptual spatial representations for indoor mobile robots."* Robotics and Autonomous Systems 56, no. 6 (2008): 493-502.
15. Hanheide, Marc, et. al *"Dora, a robot exploiting probabilistic knowledge under uncertain sensing for efficient object search."* In *Proceedings of Systems Demonstration of the 21st International Conference on Automated Planning and Scheduling (ICAPS)*, Freiburg, Germany. 2011.
16. Potapova, Ekaterina, et. al . "Attention-driven object detection and segmentation of cluttered table scenes using 2.5 d symmetry." In *Robotics and Automation (ICRA), 2014 IEEE International Conference on, pp. 4946-4952*. IEEE, 2014.
17. He, Kaiming, et. al. "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.* 2016.