

Automated Place Classification using Object Detection

Pooja Viswanathan Tristram Southey James Little
Alan Mackworth
University of British Columbia
Laboratory for Computational Intelligence
2366 Main Mall, Vancouver, British Columbia
poojav@cs.ubc.ca

Abstract

Places in an environment can be described by the objects they contain. This paper discusses the completely automated integration of object detection and place classification in a single system. We first perform automated learning of object-place relations from an online annotated database. We then train object detectors on some of the most frequently occurring objects. Finally, we use detection scores as well as learned object-place relations to perform place classification of images. Finally, we discuss areas for improvement and the application of this work to informed visual search. As a whole, the system demonstrates the automated acquisition of training data containing labeled instances (i.e. bounding boxes) and the performance of state-of-the-art object detection techniques trained on this data to perform place classification of realistic indoor scenes.

1. Introduction

The aging population has led to an increased demand for robots in the home as caretakers and assistants. The quality of life of the elderly could be dramatically improved if robots could perform simple chores such as cleaning, tidying and fetching objects. These tasks require that robots have some degree of semantic knowledge about human environments so that they can navigate, search for objects and communicate with humans successfully. For example, for a robot to perform the command “bring me an apple”, a robot must understand the term “apple” and to know which places in the environment are likely to contain it (e.g., the kitchen).

Unlike rooms that are defined by geometric properties of the environment (e.g., walls), places are defined by the objects that they contain and the set of related tasks that occur within them [13]. Place categorization thus requires the ability to recognize objects in the environment. Our embodied visual search system, Curious George, has demon-

strated a world-class ability to find query objects in controlled indoor environments by winning the robot league of the Semantic Robot Vision Challenge (SRVC) [5], an international competition to evaluate embodied object recognition systems. Curious George constructs object recognition models based on training imagery collected from the Internet, and employs a peripheral-foveal vision system to collect a visual survey of objects in a real environment. The success of our platform in the embodied object recognition scenario presents the opportunity to leverage object maps for higher-level environment understanding.

Recognized objects and their locations can be used to automatically label places in the environment through the use of annotated databases, as demonstrated by our spatial-semantic modeling system [17]. We also demonstrated that the spatial-semantic model can be used to guide a robot to more productive locations during visual search. This system, however, assumed the ability to recognize objects perfectly. In this paper, we extend our system to use real object recognition results by incorporating detector confidences during place classification.

We seek to provide a robotic system with the ability to understand and explore the environment in an automated and scalable fashion, without extensive effort from a system designer. To this end, we use the images present in the LabelMe database [12]: a free online data source which provides a large and growing amount of human-labeled visual data, much of which contains indoor scenes suitable for place labeling and object recognition. The use of Internet imagery gives the system access to training data for a nearly unlimited number of visual classes with no extra manual effort. In addition to containing object text labels, LabelMe images also contain segmentations of objects within them, which can be used to construct accurate bounding boxes. For object detection, we use a system created by Felzenszwalb et al. [6] based on mixtures of multiscale deformable part models (DPM) to perform object detection, due to its high success in the PASCAL object detection challenge.

In Section 3, we discuss the methods used to learn object-place relations, perform object detection, and classify scenes. Experiments and results are discussed in Section 4. We conclude with future directions for this research.

2. Related Work

Object classification is a highly active area of computer vision research and there are a variety of approaches that have been employed by the robotics community. Part-based models have shown themselves to be highly effective for detection of both rigid and deformable objects [4] [1]. For classifying scenes, Torralba et al. [15] and [10] use gist and SIFT features, as well as a combination of local and global features, still producing fairly poor results for several indoor scenes. Pronobis et al. [?] use a global image descriptor to allow a mobile robot to label place types. We pursue object-based scene classification since we believe that this method is more effective for indoor scenes and generalizable to a large number of previously-unseen indoor environments.

The concept of labeling areas of a 2D map, such as that captured with Simultaneous Location and Mapping (SLAM) [3], with descriptive tags has most commonly been done in topological mapping. Topological maps describe a location as a set of “places” and “connections”. Kuipers [8] proposes the Spatial Semantic Hierarchy, where space is represented at many levels that contain different degrees of detail and semantic information. Graphs have also been employed as a means of describing topological maps. In work by Ranganathan et al. [11] graph-like maps are constructed where nodes are classified using visual object recognition. Kröse et al. have developed a series of practical systems [7] [14] in which the visual similarity between images is used to cluster regions in the environment. Place labels for the clusters, however, are provided by a human through speech.

Our work is directly inspired by Vasudevan et al. [16] who label places and functional regions of the environment based on object occurrences. Their system demonstrates impressive performance in realistic environments, but uses manually labeled training images while we access an ever-growing online database containing thousands of types of objects. In addition, we incorporate detector confidence into our predictions for more robust place labelling.

3. Automated Place Labeling

We have developed a system to categorize scenes based on object detectors learned from LabelMe images. Our system is composed of four separate components. Firstly, we perform fully automated data collection from LabelMe, thus facilitating the collection of training images used to recognize a large number of object categories. We compute a

Count Model that represents the number of times an object is observed in each place type in LabelMe. We use images from LabelMe to train windowed object detectors for the most frequently occurring objects. Finally, we learn a Place Model that is used to predict the most likely place type given the detected objects in a scene.

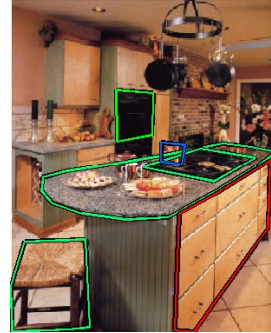


Figure 1. A kitchen scene from the LabelMe database. The polygons used to segment objects in the scene are shown as colored lines.

3.1. LabelMe Data collection

LabelMe is an online database of user annotated images. In LabelMe, the user can annotate an object in an image by selecting a region of the image using a polygon and giving that region a label. The entire scene can also have a description contained in the filename. Figure 1 shows a kitchen scene from LabelMe with several labeled objects. We use LabelMe in two ways. Training of object detectors requires tight bounding boxes that we can acquire using the LabelMe polygons. Our Place Model is learned using the correspondence between labels of objects in an image and the place name descriptor found in the image filename. Note that in creating the Place Model, we do not directly analyze the images in the dataset, and instead focus on the textual annotations in each image.

3.2. Count Model

In order to perform place classification based on objects, we first need to learn a model of objects and their number of occurrences in each place type. We can obtain this information from the LabelMe database by querying for scenes and recording the number of annotated occurrences of each object in the scene, as in Vasudevan et al. [16].

The counts table $ct_p(o)$ contains the number of times object o occurs in images of place type p . If the number of images of place type p is np , the likelihood of observing

object o in place p is computed as

$$P(o|p) = \frac{ct_p(o)}{np} \quad (1)$$

We refer to this likelihood as the Count Model, which is used to inform detector training and learning of the Place Model described below.

3.3. Detector Training

To train object detectors, we use the mixtures of multi-scale DPM described in [6]. Their system works at two levels by modeling both the entire object as well as its parts. It learns the number of parts an object consists of, the positions of these parts, and the variation in positions. The system employs a margin sensitive technique for data-mining hard negative examples to improve classification. The underlying model is called latent SVM, which is a reformulation of MI-SVM in terms of [2] latent variables. The approach alternates between fixing latent values for positive examples and optimizing the SVM object function.

The approach described above must be trained on images of the target objects with accurate bounding boxes, making many conventional data sources unusable. However, LabelMe provides user-defined bounding polygons that we use to determine training image bounding boxes. We optimized the parameters of the DPM to train all detectors in less than two hours on machines with 2 Intel quad-core Xeon 3.2 GHz processors with 32GB of memory.

We trained detectors for a subset of the most frequently occurring objects based on the Count Model. Following is a list of the objects we trained detectors for: bowl, bookshelf, cabinet, chair, cupboard, desk, dishwasher, faucet, keyboard, laptop, microwave, monitor, mouse, mousepad, mug, oven, plate, pot, refrigerator, sink, speaker, stove. The precision-recall rates for a few categories, as well as visualizations of a detector model can be found in the Experiments section.

3.4. Place Model

Given a Count Model, the Place Model is used to predict the most likely place type of the observed objects. The prior probability for each place type p is set to be uniform, since we expect to see all place types with equal probability on average in our test data. We can compute the posterior probability of the place type p given a detection det as follows:

$$P(p|det) = P(p|o)P(o|det) + P(p|\neg o)P(\neg o|det) \quad (2)$$

$P(o|det)$ represents the detection confidence as a probability derived from the SVM output. We compute this by training a sigmoid function on a hold-out set from the training

data as described in [9]. We subsequently fit SVM scores of the observed detections in the test data to this function. We compute the probability of place type p given object o as:

$$P(p|o) = \frac{P(o|p)P(p)}{\sum_i P(o|p_i)P(p_i)} \quad (3)$$

In order to predict the place type given a group of objects, we need to combine the possibly conflicting predictions of the objects present. We refer to this problem as place classification and propose the following scheme to obtain a solution. We allow every detection det observed in the test scene to contribute a vote for each place type p weighted by its posterior place probability $P(p|det)$ and the average detector precision (ap) computed during detector training. Since the detection score threshold learned during training is set to produce high recall and low precision rates, we eliminate votes of detections with very low scores. The final (unnormalized) place probability distribution (ppd) is computed as a weighted sum of the detection votes as follows:

$$ppd(p) = \sum_{det} (P(p|det) * ap_{det}) \quad (4)$$

We multiply the posterior place probability by the detector's average precision to account for the fact that some object detectors are more successful than others. The predicted label for the scene is the place type with the highest weighted sum of votes. Note that in order to achieve the best results (reported in this paper), we square the posterior probability since we have discovered empirically that while high (and low) detection scores are quite reliable, intermediate scores that lie close to the SVM margin are less reliable, and thus need to be suppressed. Future work involves investigating more theoretically founded approaches to solving the place classification problem.

4. Experiments

In this paper, we attempt to classify kitchens and offices. However, due to the automated nature of our system, we can easily learn models for other types of places including bathrooms and bedrooms by simply querying LabelMe for more scenes.

4.1. Count Model

Figure 2 shows the Count Models learned for kitchens and offices. We display the 15 most frequently occurring objects in each place type. As seen in the figures, some of the objects have unusable labels due to the ambiguous user entries. We thus select a limited number of the objects, and show later on that these are in fact sufficient for the task of place classification. Future work could involve language processing to eliminate ambiguous labels as well as combine synonymous labels together.

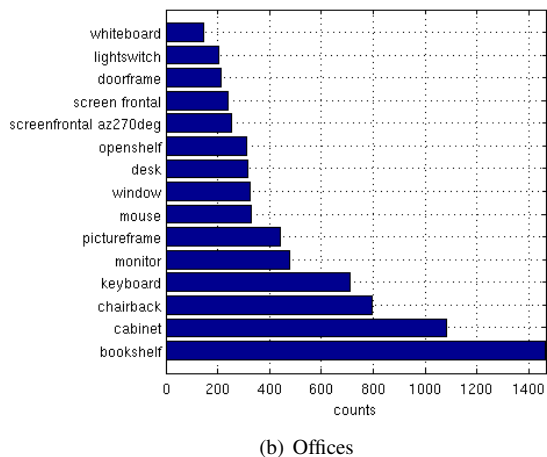
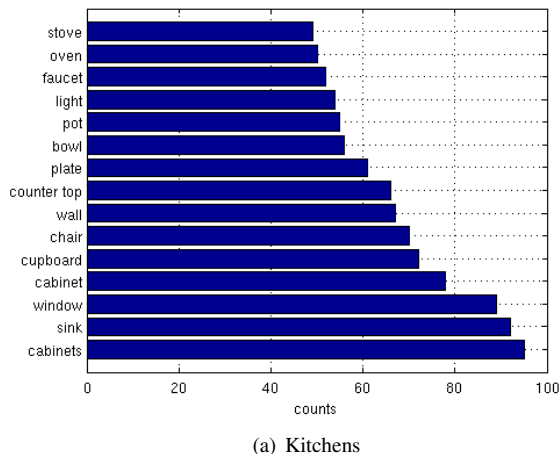


Figure 2. Counts of the types of objects found in kitchen and office scenes.

4.2. Detection

Figure 3 provides visualizations of the bowl detector using the DPM. We trained detectors using at most 200 positive examples, and 400 negative examples for each class. We set the number of components of the mixture model to 2 for most classes. Thus, training examples are split into 2 components based on the aspect ratio of the bounding boxes they contain. DPMs are trained on each component individually and merged together to form the final model. For classes with limited training data, we only used 1 component.

In order to produce precision-recall and average precision rates for each category, we validated the models on images of LabelMe objects that were not used in training. We used loosely cropped versions of these images to prevent

unannotated true positive examples in an image from being detected as false positives. Figure 4 shows some of the most and least successful detection results. As seen, objects that are usually fairly obscured by other objects (furniture such as desks and tables) tend to perform the worst. This is due to the fact that training images for these classes mostly contain views of table/desk tops. Alternate views of furniture can be gathered by using other Internet sources such as Walmart, which would provide additional structure (e.g. table legs) for use in training detectors.

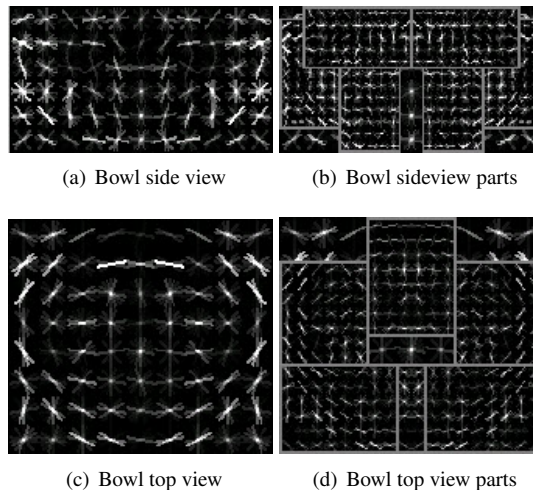


Figure 3. Visualizations of the Felzenszwalb et al. object classifier. The images on the left show an expected intensity of gradients in a grid pattern for the entire object. On the right, they show the gradients in the parts model.

4.3. Place Model

We designed experiments to test place classification in two different scenarios. In the first experiment, we attempt classification of places based on cropped images of the objects they contain. In the second, we classify full images that depict a scene containing different types and numbers of objects.

4.3.1 Place Classification of Object Images

This scenario was chosen since it most closely resembles the type of challenge faced by our robots in the SRVC challenge. In the SRVC, our robot identifies the location of unknown objects in the environment using cues such as 3D shape. Once an unknown object is found, the robot uses its

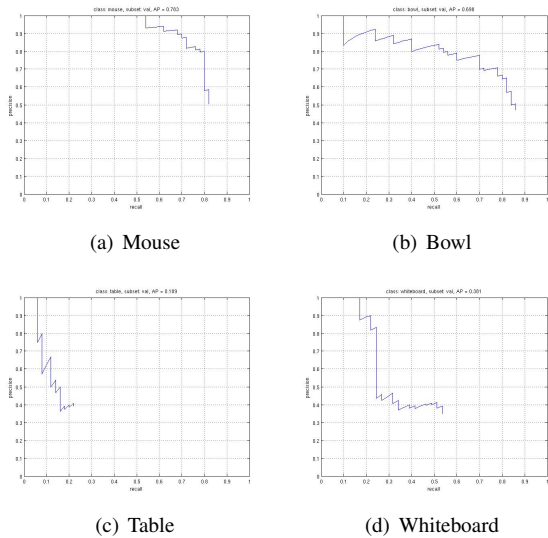


Figure 4. The precision/recall rates of object detectors. Top rows shows 3 of the most successful classifiers and the bottom row shows 3 of the least successful classifiers.

camera to take cropped high resolution images of the target object from different viewpoints.

For this experiment, we constructed a dataset of 8 home models based on real home environments in order to evaluate the place classification method. This data consists of a selection of layouts (studio apartment, regular apartment, bungalow, etc.) of varying complexity along with the locations, labels and images of some of the contained objects (see Figure 5 for an example). The images for the objects are randomly chosen cropped LabelMe images that were not used in object detection training. Additionally, for each object, we choose 5 different example images to simulate various viewpoints that the robot might collect of an object. The task is to classify the type of scene for clusters of objects that are spatially co-located using the reported detections in all object images. We assume that clustering of images based on spatial location has been performed. This can be done based on image similarity as in [7]. Alternatively, our Cluster Model described in [17] can possibly be used to perform clustering based on both spatial and semantic information. Figure 6 shows the confusion matrix for the final place labeling of object clusters. As seen, our place classification algorithm produces assignments that perfectly match the ground-truth place labels for kitchen and office scenes.



Figure 5. The layout of a house used in place classification with object images. The points represent the positions of the centroids of objects in the environment.

kitchen	8	0
office	0	7
	kitchen	office

Figure 6. The confusion matrix for place classification of object images.

4.3.2 Place Classification of Scene Images

Our second experiment is to classify a place given a single image that shows a part of the scene and can contain several objects. This is a more challenging task since the objects are not segmented and we only have a single image that does not show the entire scene. Place classification is performed by running all of our object detectors on each test image, and using the resulting detections as input to the Place Model, which infers the place label.

Images from the inside of houses were downloaded from internet websites such as Photobucket. To prevent bias from our knowledge of the types of objects we were using to classify scenes, images were labelled as either kitchen, office or other by third parties using a questionnaire. A total of 67 images were labeled by them as kitchens and offices and used for this experiment.

The results of this classification are shown as a confusion matrix in Figure 7. Given the difficulty of the task, our model performs extremely well at distinguishing between offices and kitchens based on a relatively small set of trained detectors. This demonstrates that objects present in a scene are very useful in classifying it. Future work involves testing our approach on the large database of indoor scenes used in [10], where the precision rates reported for kitchens and offices are fairly low.

kitchen	25	9
office	8	25
	kitchen	office

Figure 7. The confusion matrix for place classification of scene images.

5. Discussion

One of the greatest challenges we encounter in this work is acquiring good training data from LabelMe. For many classes, such as pot, the images in LabelMe are of different types of sub-classes (such as flower pots, ornamental pots

and cooking pots). In future work, we would like to automate clustering of objects into different types using techniques such as gist descriptors and comparing the differences of LabelMe polygons. Also, more work is needed in text processing to identify objects that are synonymous such as stoves and ovens. We also believe that improvements could be made to our object detector to be more robust to incorrectly labeled images.

We would like to attempt place classification on data collected by our robot from the SRVC competition. With our robot, we are able to acquire 3D layouts of the environment captured with a panning laser range finder. This 3D data allows us to identify structures such as tables and desks and automatically segment and acquire multiple images of objects on their surface. We can also use the place labels to guide visual search of novel objects using the Location Model described in [17]. In addition, we need to investigate the use of place labels as context to enhance recognition of objects that are currently difficult to recognize. Finally, we need to incorporate spatial relationships between objects to enhance place classification and informed search.

6. Conclusion

In conclusion, we have demonstrated a system that can perform place classification using object detection on both segmented images of objects from the environment and full scene images. In addition, we have shown that, with state-of-the-art object detectors trained with large, freely available data sources like LabelMe, we can effectively both detect and classify a wide variety of objects in realistic indoor images.

References

- [1] Y. Amit and A. Trouvé. Pop: Patchwork of parts models for object recognition. *Int. J. Comput. Vision*, 75(2):267–282, 2007.
- [2] S. Andrews, T. Hofmann, and I. Tsochantaridis. Multiple instance learning with generalized support vector machines. In *Eighteenth national conference on Artificial intelligence*, pages 943–944, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.
- [3] H. Choset and K. Nagatani. Topological simultaneous localization and mapping (slam): Toward exact localization without explicit localization. *IEEE Transactions on Robotics and Automation*, 17:125 – 137, April 2001.
- [4] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 10–17, Washington, DC, USA, 2005. IEEE Computer Society.
- [5] A. Efros and P. Rybski. Website: <http://www.semantic-robot-vision-challenge.org/>.

- [6] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.
- [7] B. Kröse, O. Booij, and Z. Zivkovic. A geometrically constrained image similarity measure for visual mapping, localization and navigation. In *Proceedings of the 3rd European Conference on Mobile Robots*, pages 168 – 174, Freiburg, Germany, 2007.
- [8] B. Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119:191 – 233, 2000.
- [9] J. C. Platt. Probabilities for sv machines. In *Advances in Large Margin Classifiers*, pages 213–238, 2000.
- [10] A. Quattoni and A. B. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009.
- [11] A. Ranganathan and F. Dellaert. Semantic modeling of places using objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2007.
- [12] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision (special issue on vision and learning)*, 77(1-3):157 – 173, 2008.
- [13] T. Southey and J. J. Little. Object discovery through motion, appearance and shape. In *AAAI Workshop on Cognitive Robotics, Technical Report WS-06-03*. AAAI Press, 2006.
- [14] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, and B. Kröse. Biron, where are you? - enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, October 2006.
- [15] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, 2003.
- [16] S. Vasudevan and R. Siegwart. Bayesian space conceptualization and place classification for semantic maps in mobile robotics. *Robotics and Autonomous Systems*, 56(6):522 – 537, June 2008.
- [17] P. Viswanathan, D. Meger, T. Southey, J. J. Little, and A. K. Mackworth. Automated spatial-semantic modeling with applications to place labeling and informed search. In *CRV*, pages 284–291. IEEE Computer Society, 2009.