Object-Graphs for Context-Aware Category Discovery

Yong Jae Lee and Kristen Grauman
University of Texas at Austin
yjlee0222@mail.utexas.edu, grauman@cs.utexas.edu

March 30, 2009

UT-AI-TR-09-2

Abstract

How can knowing about some categories help us to discover new ones in unlabeled images? Unsupervised visual category discovery is useful to mine for recurring objects without human supervision, but existing methods assume no prior information and thus tend to perform poorly for cluttered scenes with multiple objects. We propose to leverage knowledge about previously learned categories to enable more accurate discovery. We introduce a novel object-graph descriptor to encode the layout of object-level co-occurrence patterns relative to an unfamiliar region, and show that by using it to model the interaction between an image's known and unknown objects we can better detect new visual categories. Rather than mine for all categories from scratch, our method can continually identify new objects while drawing on useful cues from familiar ones. We evaluate our approach on benchmark datasets and demonstrate clear improvements in discovery over conventional purely appearance-based baselines.

1. Introduction

The goal of unsupervised visual category learning is to take a completely unlabeled collection of images and discover those appearance patterns that repeatedly occur in many examples. Often, these patterns will correspond to object categories or parts, and the resulting clusters or visual "themes" are useful to summarize the images' content, or to build new models for object recognition using little or no manual supervision [23, 3, 6, 19, 14, 1, 13, 12]. The appeal of unsupervised methods is three-fold: first, they help reveal structure in a very large image collection; second, they can greatly reduce the amount of time and effort that currently goes into annotating or tagging images; and third, they mitigate the biases that inadvertently occur when manually constructing datasets for recognition. The potential reward for attaining systems that require little or no supervision is enormous, given the vast (and ever increasing) unstructured image and video content currently available—for example in scientific databases, news photo archives, or on the Web.

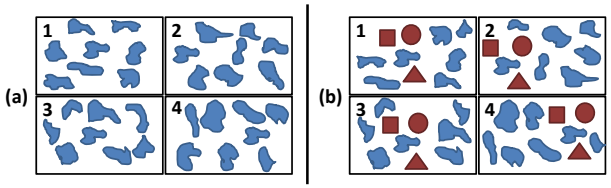


Figure 1. Toy example giving the intuition for context-aware discovery. First cover (b) and try to discover the common object(s) that appear in the images for (a). Then look at (b) and do the same. (*Hint: the new object resembles an 'r'.*) **(a)** When all regions in the unlabeled image collection are unfamiliar, the discovery task can be daunting; appearance patterns alone may be insufficient. **(b)** However, the novel visual patterns become more evident if we can leverage their relationship to things that are familiar (i.e., the circles, squares, triangles). We propose to discover visual categories within unlabeled natural images by modeling interactions between the unfamiliar regions and familiar objects.

Existing unsupervised techniques essentially mine for frequently recurring appearance patterns, typically employing a clustering algorithm to group local features across images according to their texture, color, shape, etc. Unfortunately, learning multiple visual categories simultaneously from unlabeled images remains understandably difficult, and the performance of current methods deteriorates in the presence of substantial clutter and scenes with multiple objects. While appearance is a fundamental cue for recognition, it can often be too weak of a signal to reliably detect visual themes in unlabeled, unsegmented images. In particular, appearance alone can be insufficient for discovery in the face of occluded objects, large intra-category variations,

or low-resolution data.

In this work, we propose to discover novel categories that occur amidst *known* objects within un-annotated images. How could visual discovery benefit from familiar objects? The idea is that the relative layout of understood visual objects surrounding less familiar image regions can help to detect patterns whose correct grouping may be too ambiguous if relying on appearance alone (see Figure 1). Specifically, we propose to model the interaction between a set of detected categories and the unknown to-be-discovered categories, and show how a grouping algorithm can yield more accurate discovery if it exploits both object-level context cues as well as appearance descriptors.

As the toy example in Figure 1 illustrates, novel recurring visual patterns ought to be more reliably detected in the presence of familiar objects. Studies in human perception confirm that humans use contextual cues from familiar objects to learn entirely new categories [10]. As a rough analogy for this visual process, take natural language learning: when we encounter unfamiliar words, their definition can often be inferred using the contextual meaning of the surrounding text [26].

To implement this idea, we introduce a context-aware discovery algorithm. Our method first learns category models for some set of known categories. Given a new set of completely unlabeled images, it predicts occurrences of the known classes in each image (if any), and then uses those predictions as well as the image features to mine for common visual patterns. For each image in the unlabeled input set, we generate multiple segmentations in order to obtain a pool of regions likely to contain some full objects. We classify each region as known (if it belongs to one of the learned categories) or unknown (if it does not strongly support any of the category models). We then group the unknown regions based on their appearance similarity and their relationship to the surrounding known regions. To model the inter-category interactions, we propose a novel *object-graph* descriptor that encodes the layout of the predicted classes (see Figure 2). The output of the method is a set of discovered categories—that is, a partitioning of the unfamiliar regions into coherent groups.

The proposed method strikes a useful balance between current recognition strategies at either end of the supervision spectrum. The norm for supervised image labeling methods is forced-choice classification, with the assumption that the training and test sets are comprised of objects from the same pool of categories. On the other hand, the norm for unsupervised recognition is to mine for all possible categories from scratch [23, 6, 19, 14, 1, 13, 12]. In our approach, the system need not know how to label every image region, but instead can draw on useful cues from familiar objects in order to better detect novel ones. Ultimately we envision a system that would continually expand its set of known categories—alternating between detecting what's familiar, mining among what's not, and then presenting discovered clusters to an annotator who can choose to feed the samples back as additional labeled data for new or existing categories.

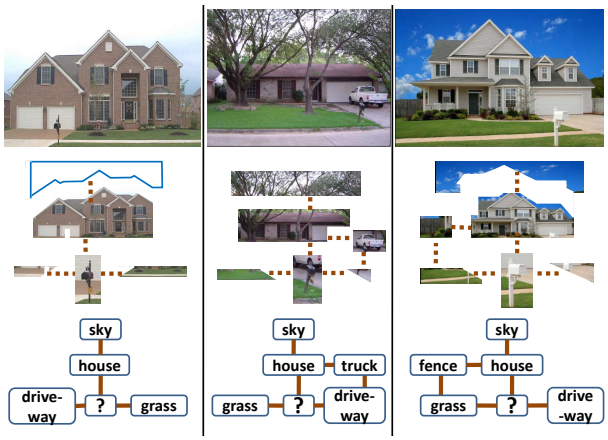


Figure 2. We would like to encode the layout of known categories relative to an unknown object. In this example, the unknown region is the *mailbox*. Our goal is to form clusters on the basis of the similarity of the unknown regions' appearance, as well as the similarity between the graphs implied by surrounding familiar objects.

Our main contribution is the idea of context-aware unsupervised visual discovery; our technique introduces (1) a method to determine whether regions from multiple segmentations are known or unknown, as well as (2) a new object-graph descriptor to encode object-level context. Unlike existing approaches, our method allows the interaction between known and unknown objects to influence the discovery. We evaluate our approach on the MSRC, Corel, and PASCAL 2008 datasets, and show that it leads to significant improvements in category discovery compared to strictly appearance-based baselines.

2. Related Work

In this section we briefly review relevant work in unsupervised category discovery and the use of context for supervised object recognition.

Existing unsupervised methods analyze appearance to discover object categories, often using bag-of-words representations and local patch features. Several methods consider models developed initially for text, such as Latent Semantic Analysis or Latent Dirichlet Allocation, to discover visual topics [23, 19, 3, 14]. Others partition the image collection using spectral clustering [6, 13, 12], or identify good exemplars with affinity propagation [1]. Our motivation is similar to these methods: to decompose large unannotated image collections into their common visual patterns or categories. However, while all previous methods

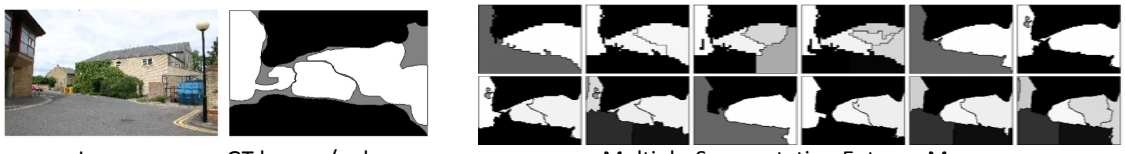


Figure 3. An example image, its ground-truth known/unknown label image, and our method's predicted entropy maps for each of its 12 segmentations. For the ground-truth, black regions denote **known** classes (sky, road), and white regions denote **unknown** classes (building, tree). (Gray pixels are "void" regions that were not labeled in the MSRC ground-truth). In the entropy maps, lighter/darker colors indicate higher/lower entropy, which signals higher/lower uncertainty according to the known category models. Note that the regions with highest uncertainty (whitest) correspond correctly to unknown objects, while those with the lowest uncertainty (darkest) are known. Regions that are comprised of both known and unknown objects are typically scored in between (gray). By considering confidence rates among multiple segmentations, we can identify the regions that are least strongly "claimed" by any known model.

assume no prior knowledge, the proposed approach allows inter-category interaction between familiar and unfamiliar regions to influence the groupings.

The idea of transferring knowledge obtained from one domain to a disjoint but similar domain is explored for object recognition in [2]; the authors devise a prior on the model parameters based on previously learned categories, thereby learning with fewer labeled examples. In contrast, we directly model the interaction *between* the learned objects and the unknown to-be-discovered objects, thereby obtaining more reliable groups from unlabeled examples.

For supervised methods that learn from labeled images, several types of context have been proposed. Global image features [24] and 3D scene layout [9] help to model the relationship between objects and their scene context. Spatial context in the form of neighboring region information can be modeled with pairwise relations [7], and with inter-pixel or inter-region spatial interactions [21, 8], or top-down constraints [22]. The benefit of high-level semantic context based on objects' co-occurrence and relative locations is demonstrated in [18, 4], and recent work shows that without such information, impoverished appearance (e.g., due to low resolution) can severely hurt recognition accuracy [17].

Our method exploits high-level semantic context for the purpose of category discovery. Unlike these supervised methods, we do not learn about inter-category interactions from a labeled training set. Instead, we identify contextual information in a data-driven manner, by detecting patterns in the relative layout of known and unknown object regions within unlabeled images.

3. Approach

The goal is to discover categories in unlabeled image collections using appearance and object-level semantic context cues. Our approach first acquires knowledge from a set of labeled "known" category examples and builds classifiers for each class. Then, given a new collection of unlabeled data, we segment each image into coherent regions. To increase the likelihood of obtaining some regions that correspond to true objects, we work with multiple segmentations. We then classify each region as "known" or "unknown" depending on the confidence that the region belongs to one of the learned categories. For each unknown region, we represent its interaction with surrounding known objects via the proposed object-graph, which encodes both the class distributions and their relative displacement. Finally, we group together regions from all images that have similar appearance and object-graphs.

Thus there are three main steps: (1) detecting instances of known objects in each image while isolating regions that are likely to be unknown; (2) extracting object-level context descriptions for the unknown regions; and (3) clustering the unfamiliar regions based on these cues. In the following, we describe each step in turn.

3.1. Identifying Unknown Objects

Any image in the unlabeled collection may contain multiple objects, and may have a mixture of familiar and unfamiliar regions. In order to describe the interaction of known and unknown objects, first we must predict which regions are likely instances of the previously learned categories[1].

Ideally, an image would first be segmented such that each region corresponds to an object; then we could classify each region and take only those with the most confident outputs as "knowns". In practice, due to the non-homogeneity of many objects' appearance, bottom-up segmentation algorithms (e.g. [20]) cannot produce such complete regions. Therefore, following [19, 16], we generate *multiple segmentations* per image, with the expectation that although some regions will fail to agree with object boundaries, some will be good segments that correspond to coherent objects. Each segmentation is the result of varying the parameters to the segmentation algorithm (i.e., number of regions, image scale).

[1]The problem of distinguishing known regions from unknown regions has not directly been addressed in the recognition literature, to our knowledge, as most methods are interested either in classifying the image as a whole, labeling every pixel with a category, or localizing a particular object.

Our idea is to compute the confidence that any of these regions correspond to a previously learned category. Assuming reliable classifiers, we will see the highest certainty for the "good" regions that are from known objects, lower responses on regions containing a mix of known and unknown objects, and the lowest certainty for regions comprised entirely of unknown objects (see Figure 3). Using this information to sort the regions, we can then determine which need to be sent to the grouping stage as candidate unknowns, and which should be used to construct the surrounding object-level cues.

We use a labeled training set to learn classifiers for N categories, $C = \{c_1, \ldots, c_N\}$. The classifiers must accept an image region as input and provide a confidence of class membership as output. We combine two appearance-based classifiers: one based on the region's appearance itself, and one that uses the region's nearby features. For the former we use a bag-of-features (BOF) with an SVM classifier; for the latter we use the discriminatively selected TextonBoost (TB) features and boosted classifier of [21]. From both classifiers we can obtain posterior probabilities for any region. We combine the outputs given by the two classifiers to compute the probability that a segment s belongs to class c_i as: $\mathrm{P}(c_i|s) = \frac{1}{2}(\mathrm{P}_{TB}(c_i|s) + \mathrm{P}_{BOF}(c_i|s))$.

The familiarity of a region is captured by the list of these posterior probabilities for each class; they reflect the class-label confidences given the region itself and its nearby appearance features. Segments that look like a learned category c_i will have a high value for $\mathrm{P}(c_i|s)$, and low values for $\mathrm{P}(c_j|s)$, $\forall j \neq i$. These are the known objects. Unknown objects will have more evenly distributed values among the posteriors. To measure the degree of uncertainty, we compute the entropy E for a segment s, $E(s) = -\sum_{i=1}^{N} \mathrm{P}(c_i|s) \cdot \log_2 \mathrm{P}(c_i|s)$.

The lower the entropy, the higher the confidence that the segment belongs to one of the known categories. Similarly, higher entropy regions have higher uncertainty and are thus more "unknown". This gives us a means to separate the known regions from the unknown regions in each image. Note that entropy ranges from 0 to $\log_2(N)$; we simply select a cutoff threshold equal to the midpoint in this range, and treat regions above the threshold as unknown and those below as known. Figure 3 shows the entropy maps we computed for the multiple segmentations from a representative example image. Note the agreement between the highest uncertainty ratings and the true object boundaries.

3.2. Object-Graphs: Modeling the Topology of Category Predictions

Given the unknown regions identified above, we would like to model their surrounding contextual information in the form of object interactions. Specifically, we want to build a graph that encodes the topology of adjacent regions relative to an unknown region (see Figure 2). Save the unknown regions, the nodes are named objects, and edges connect adjacent objects. With this representation, one could then match any two such graphs to determine how well the object-level context agreed for two candidate regions that might be grouped. Regions with similar surrounding context would have similar graphs; those with dissimilar context would generate dissimilar graphs.

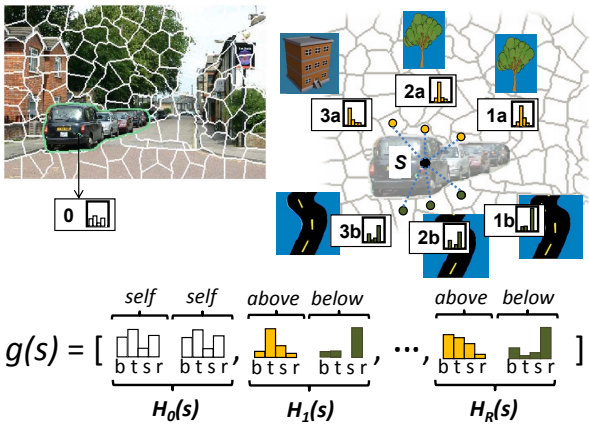


Figure 4. Schematic of the proposed object-graph descriptor. The base segment is s. The numbers indicate each region's rank order of spatial proximity to s for two orientations, *above* and *below*. The circles denote each segment's centroid. In this example, there are four known classes: building (**b**), tree (**t**), sky (**s**), and road (**r**). Each histogram $H_r(s)$ encodes the average posteriors for the r neighboring segments surrounding s from above or below, where $0 \leq r \leq R$. (Here, $R = 3$, and bars denote posterior values.) Taken together, $g(s)$ serves as a soft encoding of the likely classes that occur relative to s, from near to far, and at two orientations.

If we could rely on perfect segmentation, perfect classification, *and* perfect separation of known and unknown regions, this is exactly the kind of graph we would construct—we could simply count the number and type of known objects and record their relative layout. In practice, we are limited by the accuracy and confidence values produced by our classifier as well as the possible segments. While we cannot rectify mislabeled known/unknown regions[2], we can be more robust to misclassified known regions (e.g., sky that could almost look like water) by incorporating the uncertainty into the surrounding object context description.

We propose an *object-graph* descriptor that encodes the likely categories within the neighboring segments and their proximity to the unknown base segment. Rather than form nodes solely based on a region's class label with the maximum posterior probability, we create a histogram that forms localized counts of object presence weighted according to

[2]One might be able to avoid a hard cut and carry the entropy ratings through to the grouping stage, allowing them to influence cluster preferences; however, we have not explored this option in our implementation.

each class's posterior. For each segment, for each of two orientations (*above* and *below*) relative to its center, we compute a distribution that averages the probability values of each known class that occurs within that segment's r spatially nearest neighboring segments (where nearness is measured by distance between segment centroids), incremented over increasing values of r (see Figure 4).

For each unknown segment s, we consider $R+1$ total $2N$-dimensional histogram vectors $H_r(s)$, for $r = 0, \ldots, R$. Each histogram accumulates the average probability of occurrences of each class type c_i within s's r spatially nearest segments for each of two orientations, *above* and *below* the segment. Note that higher values of r produce a component $H_r(s)$ covering a larger region. We concatenate the R histograms to form our final object-graph descriptor for s: $g(s) = [H_0(s), \ldots, H_R(s)]$. The result is an $R \cdot 2N$-dimensional descriptor that softly encodes the surrounding objects present in increasingly further spatial extents.

We select a value of R large enough to typically include all surrounding regions in the image. We limit the orientations to above and below (as opposed to also using left and right) since we expect this relative placement to have more semantic significance; objects that appear side-by-side can often be interchanged from left-to-right (e.g., see the mailbox example in Figure 2).

For images that contain multiple unknown objects, we do not exclude the class-probability distributions of the unknown regions present in another unknown region's object-graph. Even though the probabilities are specific to known objects, their distributions still give information about appearance and surroundings of unknown objects. That is, although the probabilities cannot denote which class the unknown region should belong to (since all possible answers would be incorrect), they will still produce similar distributions for similar-looking unknown regions. As long as the unknown objects consistently appear in similar surrounding displacements throughout the dataset (e.g, unfamiliar cows appearing near other unfamiliar cows), it should only aid the contextual description.

Previous methods have been proposed to encode the appearance of nearby regions or patches [21, 8, 13, 25], however our object-graph is unique in that it describes the neighborhood of a region based on object-level information, and explicitly reflects the layout of previously learned categories. (In Section 4 we demonstrate the comparative value for the discovery task.) Relative to existing graph kernels from the machine learning literature [5, 11], our approach allows us to represent object topology without requiring hard decisions on object names and idealized segmentations.

3.3. Category Discovery Amidst Familiar Objects

In order to discover categories, we need to form homogeneous groups from the collection of unknown regions such that each group contains a number of regions with similar appearance and surrounding object context. If at least one segment in the multiple segmentations for each image corresponds to an actual unknown object, then good matches can be made among those that belong to the same category.

We define a similarity function between two regions s_m and s_n that includes both region appearance and known-object context. To describe appearance, we use a BOF representation. To describe *known-object context*, we use our object-graph descriptor. We compare both using χ^2 kernels. The final affinity between segments is the combined kernel values:

$$K(s_m, s_n) = \frac{1}{2}\left(K_{app}(s_m, s_n) + K_{obj-graph}(s_m, s_n)\right).$$

We compute affinities between all pairs of unknown regions to generate an affinity matrix. This matrix is then input to a spectral clustering algorithm to group the regions. We use the method developed in [15].

Since our unknown/known separation for novel images may be imperfect, some discovered groups may contain objects that actually belong to a known class. To rank the segments within a group according to their intra-cluster similarity, we sort them based on their degree: $D(s_m) = \sum_{l \in L} K(s_m, s_l)$, where L denotes the cluster containing segment s_m. This degree reflects how "central" a segment is to the discovered cluster. (This is how example regions are chosen for display in Section 4.)

4. Results

To validate our approach, we use the MSRC-v2 [21], PASCAL 2008, and Corel datasets. The MSRC contains 21-classes and 591 images, the PASCAL contains 20-classes and 1023 images (we use the trainval set from the segmentation tester challenge), and the Corel contains 7-classes and 100 images. Our dataset selection was based on the requirement that we have (1) images with pixel-level ground truth and (2) images with multiple objects from multiple categories that would allow us to model category interactions. To our knowledge, these are the best and most recent sets satisfying these requirements. We evaluate all sets for accuracy, and focus additional analysis on the MSRC since it has the largest number of categories, and ground-truth labeling for all objects in the dataset.

For the MSRC, PASCAL, and Corel, we chose {*building, tree, cow, sheep, bicycle*}, {*bus, diningtable, dog, horse, motorbike, tvmonitor*}, and {*rhino/hippo, polarbear*}, respectively, as unknown classes[3]; these had the

[3]The known classes for the MSRC, PASCAL, and Corel are {*grass, sky, airplane, water, face, car, flower, sign, bird, book, chair, road, cat,*

most inter-category interactions. For each dataset, we learn the "known" classes on 55% of the data and run our discovery algorithm on the other 45%.

Implementation Details: In all experiments, we use an RBF kernel wrapped around χ^2 distances, with $\gamma = 1$; we set this parameter on the validation sets. We use Normalized Cuts [20] to generate multiple segmentations for each image by varying the number of segments, $M = 3, 5, 7, 9$, and applying these settings at three image scales: 50, 100, and 150 pixels across, following [19]. This results in 12 segmentations (72 segments) per image.

The BOF representation that we use to describe region appearance for classification and clustering are the Bag-of-Words (BOW), Texton Histogram (TH), and Color Histogram (CH). For the BOW, we quantize densely extracted SIFT descriptors (every 10 pixels in the image, radius-8 patches) to 1000 visual words. We compute the TH feature by convolving each image with 18 bar and 18 edge filters (6 orientations and 3 scales for each), 1 Gaussian, and 1 Laplacian-of-Gaussian filters. These responses are quantized to 400 texton words. For BOW and TH, we form the vocabulary using k-means. We compute CH features in Lab color space, with 23 bins for the L, a, and b, respectively.

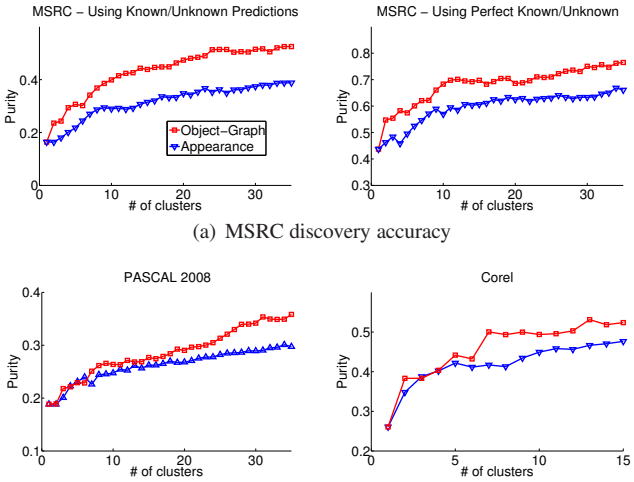
For the PASCAL, we use BOW, and for the Corel, we concatenate TH and CH features. For the MSRC, we use BOW for classification and TH for clustering. These feature choices follow previous related work on these datasets [21, 18, 7]. To compute class probabilities for the regions we use one-vs-one SVM classifiers, and map their outputs to multi-class posteriors using the pairwise coupling approach of [27]. For our object-graph descriptor, we fix the neighborhood range at $R = 20$ per orientation.

Evaluation Metric: We use *purity* to quantify the accuracy of our method's category discovery. Purity measures the extent to which a cluster contains regions of a single dominant class. Since an output region may not agree entirely with true object boundaries, we evaluate it according to the label of the majority pixel-level ground-truth label. We only consider regions with ground-truth labels for clustering. We vary the number of clusters and check purity for each value, since we cannot assume prior knowledge on the number of novel categories. Since the spectral clustering step [15] uses a random initialization, we perform 10 runs for each experiment and average the results.

4.1. Unsupervised Discovery Accuracy

To support our claim that the detection of familiar objects should aid in unsupervised category discovery, we need to evaluate how much accuracy improves when we form groups using appearance together with the object-graph, versus when we form groups using appearance alone. We thus generate two separate curves for purity scores: (1) an appearance-only baseline where we cluster unknown regions using only appearance features (Appearance), and (2) our approach, where we cluster using both appearance and contextual information (Object-Graph).

dog, body, boat}, {*airplane, bicycle, bird, boat, bottle, car, cat, chair, cow, person, pottedplant, sheep, sofa, train*}, and {*water, snow, vegetation, ground, sky*}, respectively.



(a) MSRC discovery accuracy

(b) PASCAL08 and Corel discovery accuracy

Figure 5. Purity rates for the MSRC (top) and PASCAL 2008 and Corel images (bottom). Higher curves are better. We compare our approach (Object-Graph) with an appearance-only (Appearance) baseline. The discovered categories are more accurate using the proposed approach, as the familiar objects nearby help us to detect region similarity even when their appearance features may only partially agree. The top right plot shows the performance attainable were we to perfectly separate segments according to whether they are known or unknown (see text).

Since our evaluation scenario necessarily differs from earlier work in unsupervised discovery (we assume some background knowledge about a subset of the dataset's classes), it is not possible to meaningfully compare the output of our method with previously reported numbers. However, the appearance-only baseline is intended as a broad stand-in for previous unsupervised methods, which all rely solely on appearance.

Figure 5 shows the results for all three datasets. Our model significantly outperforms the appearance-based approach. These results confirm that the appearance and object-level contextual information complement each other to produce high quality clusters. The gains for the MSRC and Corel are most significant; on the PASCAL 2008 data we see somewhat narrower improvements, most likely due to the greater difficulty of the data, but also due to the limited amount of repeated objects for the known context.

The top right plot in the same figure shows the results for the MSRC if we replace our known-unknown predictions with perfect separation (note the vertical axis scale change). Again our model significantly outperforms the appearance-

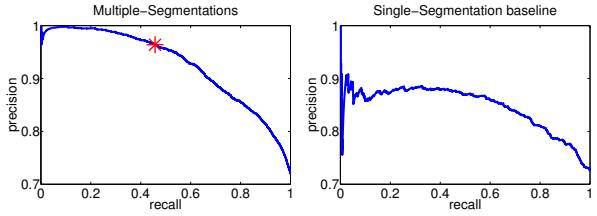


Figure 6. Precision / Recall curves for known vs. unknown decisions. Our cutoff threshold is set to half of the maximum entropy value, and the Precision / Recall value at that point is indicated by the red star. We compare against a single-segmentation baseline. These results indicate that our multiple-segmentations approach provides better estimates of known and unknown regions, which can lead to better discovery.

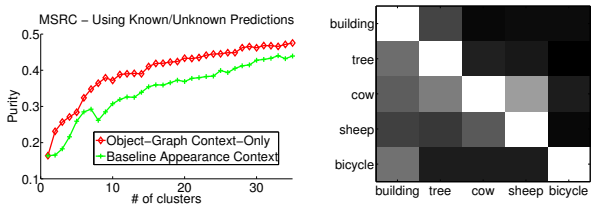


Figure 7. (**left**) We compare context-only cluster rates computed using our Object-Graph descriptor with those from a baseline descriptor on the MSRC. Our class-based descriptor outperforms the appearance-based context baseline. (**right**) Confusion matrix showing which categories are most often mistakenly grouped together during discovery.

only baseline. The purity rates are notably higher here compared to when the known/unknown separation is computed automatically based on entropy scores. The reason is that our discovery problem has become much simpler: instead of having image segments that could belong to one of 21 categories (total number of known and unknown categories), we only need to consider the 5 unknown categories. This implies that there is room for better initial classification (i.e., better label predictions and confidences for the known and unknown regions), with which we can expect higher cluster rates. Currently our known-category classifier correctly labels segments 78.55% of the time on the MSRC.

Figure 6 shows the precision/recall curves for our known-unknown decisions on the MSRC. For this, we set the known classes as positive, and the unknown classes as negative, and sort the regions by their entropy scores in ascending order. We compare against a single-segmentation baseline where each image is segmented into eight regions. The curve falls more quickly for the single-segmentation case, indicating that confusion between known and unknown regions occurs earlier on. With multiple-segmentations, the confusion occurs much later.

The red star indicates the precision/recall value at $\frac{1}{2}\max E(s)$. With this threshold, we accurately consider almost all of the unknown regions for discovery. At the same time, those that are not considered are almost always known regions. However, there are also many known regions that are incorrectly identified as unknown. While this makes the problem of discovery more difficult for the unknown objects, it also allows "re-discovery" of known categories.

4.2. Impact of the Object-Graph Descriptor

We next evaluate how our object-graph descriptor compares to a simpler alternative that directly encodes the surrounding appearance features. Since part of our descriptor's novelty rests on its use of object-level information, this is an important distinction to study empirically. To do this, we substitute class probability counts for the object-graph with TH histogram counts. That is, instead of describing context with estimates of known-category objects, we look directly at the surrounding appearance information given by the TH features. Figure 7 (left) shows purity rates computed by our object-graph descriptor and the baseline descriptor. Our object-graph performs noticeably better than the baseline. This confirms that directly modeling class-interactions instead of surrounding appearance information improves discovery. Even though our initial classification results are based on the same information, learned category distributions are more reliable than local appearance patterns.

In addition to improved accuracy, our descriptor also has the advantage of lower dimensionality. The object-graph requires only $R \cdot 2N$-dimensional vectors for each unknown region, whereas the TH baseline requires $R \cdot 2Q$-dimensional vectors, for Q texton words, and N known classes. In this case, $N = 16$ and $Q = 400$, so our object-graph is 25 times more compact.

4.3. Discovered Categories and Confusions

Figure 8 shows examples of discovered categories using our approach and the appearance-only baseline. For this experiment, we set the number of clusters, $k = 21$, equal to the number of classes. The cluster images are sorted by their degree (top left is highest, bottom right is lowest). We remove overlapping regions and show only one region per image. We show the top ten images for each cluster.

Finally, we compute a confusion matrix for our discovered regions by counting co-occurrence for pairs of categories across the clusters. Specifically, for each category c_i, we count how many segments of that category are clustered together with segments belonging to category c_j. We normalize the result by the total number of segments belonging to c_i to account for the variability in the number of segments per category. Figure 7 (right) shows the result for the five unknown categories of the MSRC. The highest confusion occurs between the *cow* and *sheep* classes, which is reasonable considering they have similar surroundings (i.e., *grass* and *sky*). Learning the weight parameters on the context

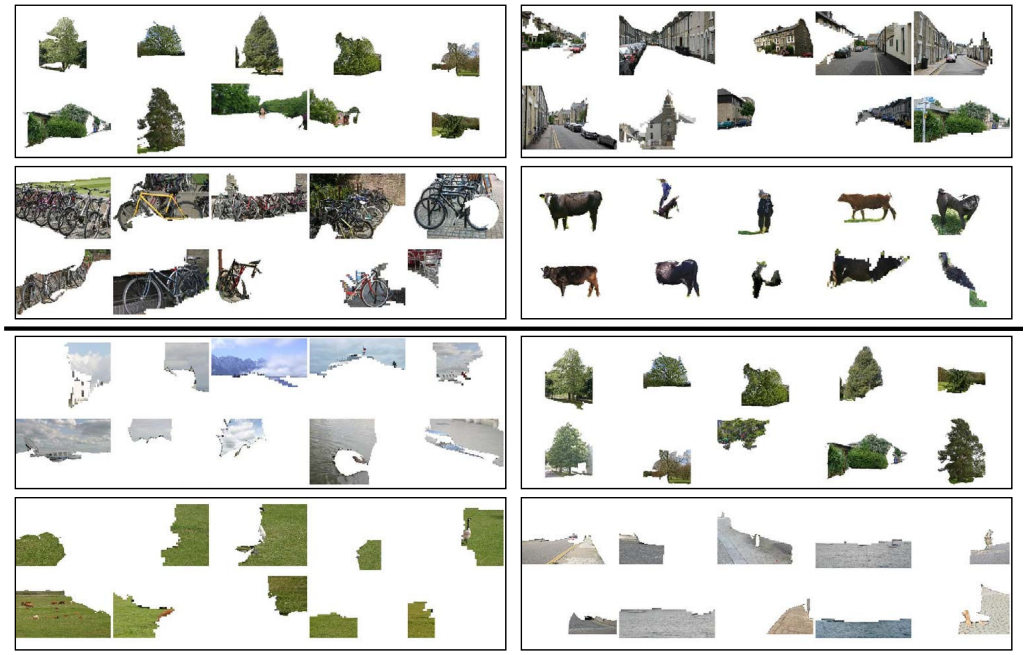


Figure 8. Examples of discovered categories for the MSRC. The images are sorted from top left to bottom right by how "central" they are to the cluster (see text for details). The first two rows correspond to discovered categories *Tree, Building, Bicycle,* and *Cow*. The last two rows correspond to some discovered categories when we set as unknown *Grass, Sky, Road, Water,* and *Tree*. The groups are quite consistent with semantic categories, and often are more inclusive than those that would be found with appearance alone.

and appearance kernels could potentially reduce such confusions (we have simply fixed them to be weighted equally).

5. Conclusions

We have developed an algorithm that models the interaction between familiar categories and unknown regions to discover novel categories in unlabeled images. We believe that our system could be used in a loop where an annotator could label the meaningful discovered clusters, which would then become the familiar objects for which a classifier can be trained. This would expand the object-level context for future discovery and continually increase the number of known categories.

References

[1] D. Dueck and B. Frey. Non-metric Affinity Propagation for Unsupervised Image Categorization. In *ICCV*, 2007. 1, 2

[2] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories. In *ICCV*, 2003. 3

[3] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. In *ICCV*, 2005. 1, 2

[4] C. Galleguillos, A. Rabinovich, and S. Belongie. Object Categorization using Co-Occurrence, Location and Appearance. In *CVPR*, 2008. 3

[5] T. Gartner, P. Flach, and S. Wrobel. On Graph Kernels: Hardness Results and Efficient Alternatives. In *COLT*, 2003. 5

[6] K. Grauman and T. Darrell. Unsupervised Learning of Categories from Sets of Partially Matching Image Features. In *CVPR*, 2006. 1, 2

[7] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale Conditional Random Fields for Image Labeling. In *CVPR*, 2004. 3, 6

[8] G. Heitz and D. Koller. Learning Spatial Context: Using Stuff to Find Things. In *ECCV*, 2008. 3, 5

[9] D. Hoiem, A. A. Efros, and M. Hebert. Putting Objects in Perspective. In *CVPR*, 2006. 3

[10] A. S. Kaplan and G. L. Murphy. The Acquisition of Category Structure in Unsupervised Learning. *Memory & Cognition*, 27:699–712, 1999. 2

[11] H. Kashima, K. Tsuda, and A. Inokuchi. Kernels on graphs. *Kernels and Bioinformatics*, 2004. 5

[12] G. Kim, C. Faloutsos, and M. Hebert. Unsupervised Modeling of Object Categories Using Link Analysis Techniques. In *CVPR*, 2008. 1, 2

[13] Y. J. Lee and K. Grauman. Foreground Focus: Unsupervised Learning From Partially Matching Images. In *BMVC*, 2008. 1, 2, 5

[14] D. Liu and T. Chen. Unsupervised Image Categorization and Object Localization using Topic Models and Correspondences between Images. In *ICCV*, 2007. 1, 2

[15] A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NIPS*, 2001. 5, 6

[16] C. Pantofaru, C. Schmid, and M. Hebert. Object Recognition by Integrating Multiple Image Segmentations. In *ECCV*, 2008. 3

[17] D. Parikh, C. L. Zitnick, and T. Chen. From Appearance to Context-Based Recognition: Dense Labeling in Small Images. In *CVPR*, 2008. 3

[18] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *ICCV*, 2007. 3, 6

[19] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In *CVPR*, 2006. 1, 2, 3, 6

[20] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *TPAMI*, 22(8):888–905, August 2000. 3, 6

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. In *ECCV*, 2006. 3, 4, 5, 6

[22] A. Singhal, J. Luo, and W. Zhu. Probabilistic Spatial Context Models for Scene Content Understanding. In *CVPR*, 2003. 3

[23] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering Object Categories in Image Collections. In *ICCV*, 2005. 1, 2

[24] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 53(2), 2003. 3

[25] A. Vedaldi and S. Soatto. Relaxed Matching Kernels for Object Recognition. In *CVPR*, 2008. 5

[26] R. Weischedel. Adaptive Natural Language Processing. In *Association for Computational Linguistics*, 1990. 2

[27] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability Estimates for Multi-Class Classification by Pairwise Coupling. *JMLR*, 5:975–1005, August 2004. 6