To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June, 2010.

# 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

## 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 identifies 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 minimal manual supervision [8, 23, 17, 14, 16]. 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 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.

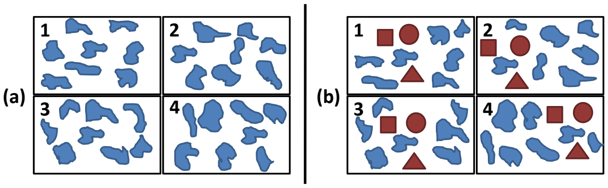

Figure 1. Toy example giving the intuition for context-aware discovery. First cover (b) and try to discover the common object(s) that appear in the images for (a). Then look at (b) and do the same. (*Hint: the new object resembles an 'r'.*) **(a)** When all regions in the unlabeled image collection are 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, especially 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 familiar 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

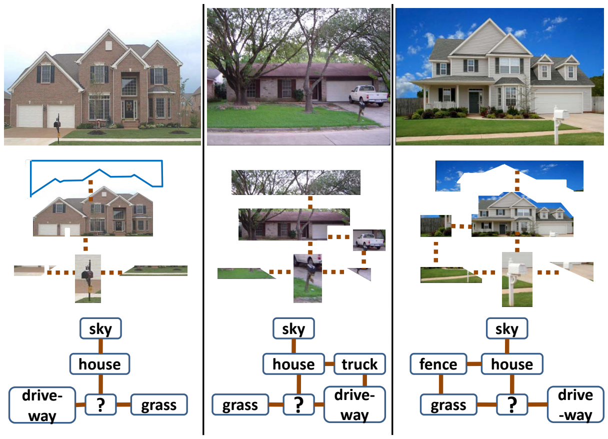

Figure 2. We want to encode the layout of known categories relative to an unknown object. In this example, the unknown region is the *mailbox*. Our goal is to form clusters on the basis of the similarity of the unknown regions' appearance, as well as the similarity between the graphs implied by surrounding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

the presence of familiar objects. Studies in perception confirm that humans use contextual cues from familiar objects to learn entirely new categories [12]. The use of familiar things as context applies even for non-vision tasks. For example, take natural language learning: when we encounter unfamiliar words, their definition can often be inferred using the contextual meaning of the surrounding text [30].

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 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, 17, 8, 14, 16]. In our approach, the system need not know how to label every image region, but instead can draw on useful cues from familiar objects 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.

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 four datasets, and show that it leads to significant improvements in category discovery compared to strictly appearance-based baselines.

## 2. Related Work

Existing unsupervised methods analyze appearance to discover object categories, often using bag-of-words representations and local patch features. Some methods leverage topic models, such as Latent Semantic Analysis, to discover visual themes [23, 17]. Others partition the image collection using spectral clustering [8, 14, 16]. 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 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 [4, 2]; the authors devise a prior 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 [27] and 3D scene layout [11] help to model the relationship between objects and scenes. Spatial context can be modeled with neighboring inter-region interactions [9, 25, 10, 18]. The benefit of high-level semantic context based on objects' co-occurrence and relative locations has also been demonstrated [5, 21, 28, 7].

Our method exploits high-level semantic context for unsupervised category discovery. Unlike the above supervised methods, we do not learn about inter-category interactions from a labeled training set, nor do we aim to improve the detection of familiar objects via context relationships. Instead, we identify contextual information in a data-driven manner,

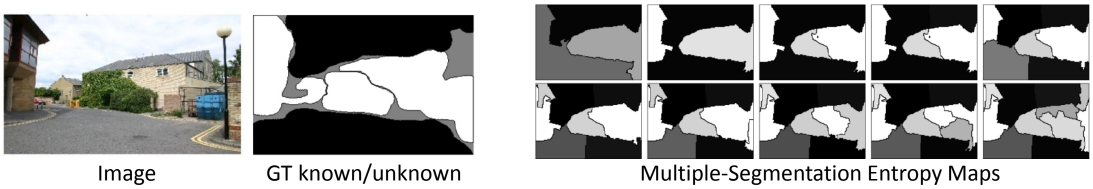

Figure 3. An example image, its ground-truth known/unknown label image, and our method's predicted entropy maps for each of its 10 segmentations. For the ground-truth, black regions denote **known** classes (sky, road), and white regions denote **unknown** classes (building, tree). (Gray pixels are "void" regions that were not labeled in the MSRC-v2 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.

by detecting patterns in the relative layout of known and unknown object regions within unlabeled images. The method in [15] recovers contextual information on-the-fly from the test images by exploiting the data's statistical redundancy. However, in contrast to our approach, that method learns the context surrounding familiar object instances to improve their classification, whereas our approach discovers object-level context surrounding *unfamiliar* object regions to improve their grouping (discovery of new objects).

## 3. Approach

There are three main steps to our approach: (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. Normalized Cuts [24]) cannot produce such complete regions. Therefore, following [23], 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). As in previous work, each segment goes into the pool of instances that will be processed by the algorithm, which means segments that overlap in the same original image are treated as separate instances.

We first 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 texture, color, and shape features using the multiple kernel learning (MKL) framework of [1] and obtain posterior probabilities for any region with an SVM classifier; i.e., the probability that a segment $s$ belongs to class $c_i$, $\mathrm{P}(c_i|s)$. (Details on the features we use in our results are given in Section 4.)

The familiarity of a region is captured by the list of these posterior probabilities for each class. 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; correspondingly, we consider a region with a high entropy score to be a likely "unknown". This gives us a means to separate each image into known and unknown regions. 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 entropy maps for the multiple segmentations from a representative example image. Note the agreement between the highest uncertainty ratings and the true object boundaries.

[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 aim to either classify the image as a whole, label every pixel with a category, or localize a particular object.

## 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.

If we could rely on perfect segmentation, classification, *and* 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, 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 each class's posterior. For each segment, 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).

Specifically, for each unknown segment $s$, we compute a series of histograms using the posteriors computed within its neighboring superpixels. Each component histogram $H_r(s)$ 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. We concatenate the component histograms for $r = 0, \ldots, R$ to produce the final object-graph descriptor:

$$g(s) = [H_0(s), H_1(s), \ldots, H_R(s)], \qquad (1)$$

where $H_0(s)$ contains the posteriors computed within $s$'s central superpixel. The result is an $((R+1) \cdot 2N)$-dimensional vector, where $N$ denotes the number of familiar classes. Note that higher values of $r$ produce a component $H_r(s)$ covering a larger region, and the descriptor

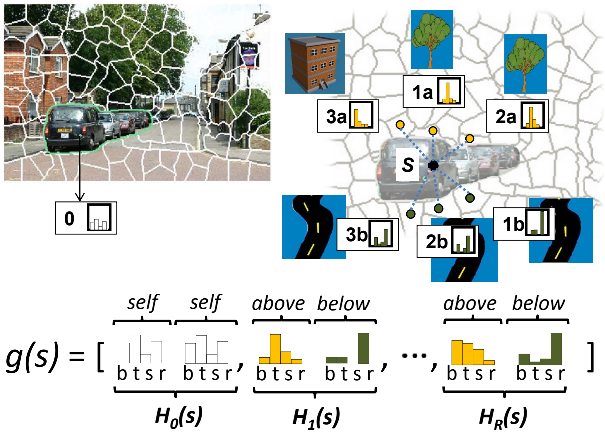

Figure 4. Schematic of the proposed object-graph descriptor. The base segment is $s$. The numbers indicate each region's rank order of spatial proximity to $s$ for two orientations, *above* and *below*. The circles denote each segment's centroid. In this example, there are four known classes: building (**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.

softly encodes the surrounding objects present in increasingly further spatial extents. Our representation can detect partial context matches (i.e., partially agreeing spatial layouts), since the matching score between two regions is proportional to how much their context agrees. Due to the cumulative construction, discrepancies in more distant regions have less influence.

There are a couple of implementation details that will help ensure that similar object topologies produce similar object-graph descriptors. First, we need to maintain consistency in the size and relative displacement of nodes (regions) across different object-graphs; to do this, we use superpixel segments as nodes (typically about 50 per image). Their fairly regular size and shape tessellates the image surrounding the unknown region well, which in turn makes a centroid-based distance between nodes reliable.[2] As usual, the superpixels may break non-homogeneous objects into multiple regions, but as long as the oversegmentation effect is fairly consistent in different images (e.g., the dark roof and light wall on the building are often in different superpixels), the object-graph will avoid misleading double-counting effects. Empirically, we have observed that this consistency holds.

Second, we need to obtain robust estimates of the known objects' posterior probabilities, and avoid predicting class memberships on regions that are too local (small). For this we exploit the multiple segmentations: we estimate the

[2] Note that our descriptor assumes images have similar scene depth, and thus that the relative placement of surrounding objects depends only on the scale of the object under consideration (as do most existing recognition methods using object co-occurrence context, e.g. [25, 10]).

class posteriors for each segment, then for each image, we stack its segmentation maps, and compute a per-pixel average for each of the $N$ posterior probabilities. Finally, we compute the posteriors for each superpixel node by averaging the $N$-vector of probabilities attached to each of its pixels. Note that this allows us to estimate the known classes' presence from larger regions, but then summarize the results in the smaller superpixel nodes.

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 weak information about the appearance of unknown objects. The probabilities cannot denote which class the unknown region should belong to (since all possible answers would be incorrect), but we will get 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 [25, 10, 29, 16], however our object-graph is unique in that it describes the region neighborhood 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 [6, 13], 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

Now that we have a means to compute object-level context, we can combine this information with region-based appearance to form homogeneous groups from our collection of unknown regions. We define a similarity function between two regions $s_m$ and $s_n$ that includes both region appearance and known-object context:

$$K(s_m, s_n) = \frac{1}{|\mathrm{u}|} \sum_{\mathrm{u}} K_{\chi^2}\left(a_{\mathrm{u}}(s_m), a_{\mathrm{u}}(s_n)\right) + K_{\chi^2}\left(g(s_m), g(s_n)\right),$$

where $g(s_m)$ and $g(s_n)$ are the object-graph descriptors as defined in Eqn. 1, and each $a_{\mathrm{u}}(s_m)$ and $a_{\mathrm{u}}(s_n)$ denotes an appearance-based feature histogram extracted from the respective region (which will be defined in Section 4). Each $K_{\chi^2}(\cdot,\cdot)$ denotes a $\chi^2$ kernel function for two histogram inputs: $K_{\chi^2}(x, y) = \exp(-\frac{1}{2}\sum_i \frac{(x_i - y_i)^2}{x_i + y_i})$, where $i$ indexes the histogram bins.

```
Input: Set of classifiers for $N$ known category models, set of novel
        unlabeled images, and $k$.
Output: Set of $k$ discovered categories (clusters).
1. Obtain multiple segmentations for each image.
2. Compute posteriors for each region. (Sec. 3.1)
3. Compute the entropy for each region to classify as "known" or
"unknown". (Sec. 3.1)
4. Construct an object-graph for each unknown region. (Sec. 3.2)
5. Compute affinities between unknown regions with the
object-graph and appearance features, and cluster to discover
categories. (Sec. 3.3)
```

**Algorithm 1**: The context-aware discovery algorithm

We compute affinities between all pairs of unknown regions to generate an affinity matrix, which is then given as input to a clustering algorithm to group the regions. We use the spectral clustering method developed in [20]. Because we use multiple segmentations, if at least one "good" segment of an unknown object comes out of an image, then it may be matched and clustered with others that belong to the same category. Since our unknown/known separation for novel images may be imperfect, some discovered groups may contain objects that actually belong to a known class. Importantly, since affinity can be boosted by either similar appearance *or* similar context of known objects, we expect to be able to discover objects with more diverse appearance.

We summarize the steps of our algorithm in Alg. 1.

## 4. Results

In this section, we (1) evaluate our method's discovery performance and compare against two appearance-only baselines, (2) analyze our entropy-based known-unknown separation measure, and (3) compare the object-graph with an appearance-based context baseline.

We validate our approach with four datasets: MSRC-v0, MSRC-v2, PASCAL VOC 2008, and Corel. The MSRC-v0 has 21-classes (3,457 images), the MSRC-v2 has 21-classes (591 images), the PASCAL has 20-classes (1,023 images; we use the trainval set from the segmentation challenge), and the Corel has 7-classes (100 images). Our dataset selection is based on the requirements that the images have pixel-level ground truth and multiple objects from multiple categories. We evaluate on all sets, and focus additional analysis on the MSRC-v2 since it has the largest number of categories, and ground-truth labeling [19] for all objects.

We want to evaluate how sensitive our method is w.r.t. which classes are considered familiar (or unfamiliar), and how many (or few) objects are in the "known" set of models. Thus for each dataset, we form multiple splits of known/unknown classes, for multiple settings of both the number of knowns ($N$) and the number of true unknowns present. Please see the supplementary file for a detailed

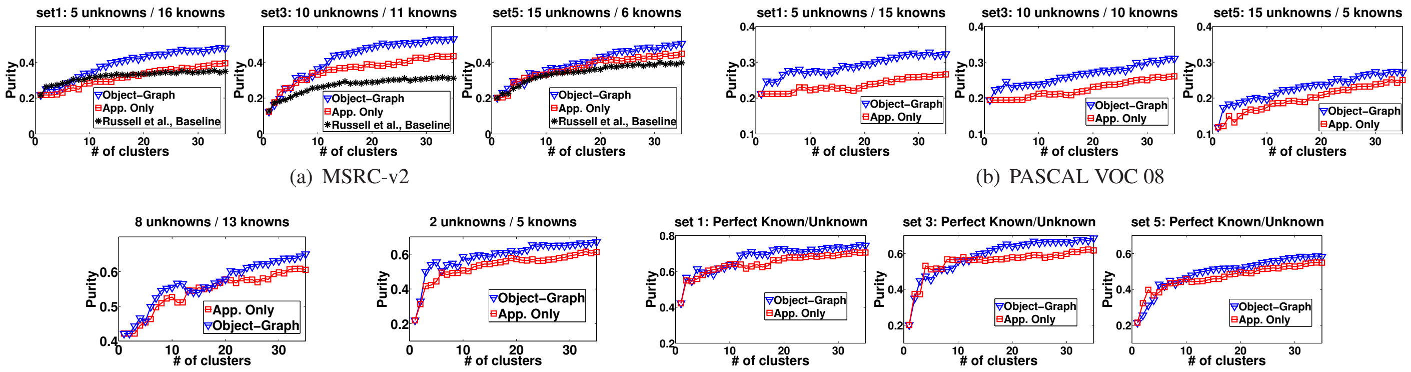

Figure 5. Discovery accuracy results. **(a) through (d):** Purity rates for all four datasets as a function of $k$. Higher curves are better. We compare our approach (Object-Graph) with appearance-only baselines. The discovered categories are more accurate using the proposed approach, as the familiar objects nearby help us to detect region similarity even when their appearance features may only partially agree. **(e):** Performance attainable were we able to perfectly separate segments according to whether they are known or unknown.

breakdown of the category names in each split. We learn the known classes on 60% of the data and run our discovery algorithm on the other 40%.

**Implementation Details:** We use Normalized Cuts [24] for segmentation, and vary the number of segments from 3 to 12 to obtain 10 segmentations (75 segments) per image. To form each appearance descriptor $a_{\mathrm{u}}(s)$ for a region $s$, we use several types of bag-of-features histograms: Texton Histograms (TH), Color Histograms (CH), and pyramid of HOG (pHOG) [3]. For TH, we use a filter bank with 18 bar and edge filters (6 orientations and 3 scales for each), 1 Gaussian, and 1 Laplacian-of-Gaussian filters. We quantize to 400 textons via $k$-means. For CH, we use Lab color space, with 23 bins per channel. For pHOG, we use 3 pyramid levels with 8 bins. We normalize each $a_{\mathrm{u}}(s)$ and $g(s)$ to sum to 1. To compute class probabilities, we use one-vs-all SVM classifiers trained using MKL, and obtain posteriors using [22]. For the object-graphs, we generate an over-segmentation with roughly 50 superpixels per image, and fix $R = 20$.

**Evaluation Metrics:** We use both *purity* [26] and *mean Average Precision* (mAP) to quantify accuracy. The former rates the coherency of the clusters discovered, while the latter reflects how well we have captured the affinities between intra-class versus inter-class instances (independent of the clustering algorithm). We only consider regions with ground-truth labels (i.e., no "voids" from MSRC). To score an arbitrary segment, we consider its ground truth label to be that which the majority of its pixels belong to.

These metrics reward discovery of object parts as well as full objects (e.g., we would get credit for discovering cow heads and cow legs as separate entities). This seems reasonable for the unsupervised category discovery problem setting, given that the part/object division is inherently ambiguous without external human supervision. We report purity values as a function of the number of clusters, since we cannot assume prior knowledge on the number of novel categories. Since the spectral clustering step [20] uses a random initialization, we average all results over 10 runs.

**Unsupervised Discovery Accuracy:** To support our claim that the detection of familiar objects should aid in category discovery, we 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 (App. only), and (2) our approach, where we cluster using both appearance and contextual information (Object-Graph).

Since our evaluation scenario necessarily differs from earlier work in unsupervised discovery, it is not possible to directly compare the output of our method with previously reported numbers: our method assumes some background knowledge about a subset of the classes, whereas existing discovery methods assume none. However, our appearance-only baseline is intended to show the limits of what can be discovered using conventional approaches for this data, since previous unsupervised methods all rely solely on appearance [23, 8, 14, 16]. Furthermore, we also generate comparisons with the state-of-the-art LDA-based discovery method of Russell et al. [23] using the authors' publicly available code. To our knowledge, theirs is the only other current unsupervised method that tests with datasets containing multiple objects per image, making it the most suitable method for comparison. In all results, our method and the baselines are applied to the same pool of segments (i.e., those our method identifies as unknown).

Figure 5 (a-d) shows the results for all of the datasets. Our model significantly outperforms the appearance-only baselines. These results confirm that the appearance and object-level contextual information complement each other

|  | Building | Tree | Cow | Airplane | Bicycle |
|---|---|---|---|---|---|
| Our full model | **0.32** | **0.36** | **0.41** | **0.36** | 0.21 |
| App. only | 0.27 | 0.33 | 0.20 | 0.21 | 0.10 |
| Obj-Graph only | **0.32** | 0.27 | 0.37 | 0.32 | **0.24** |

Table 1. Mean Average Precision (mAP) on MSRC-v2 set1 unknowns.

to produce high quality clusters.[3] Parts (a) and (b) illustrate our method's consistency with respect to various random splits of unknown/known category pools.

To directly evaluate how accurately our object-graph affinities compare the regions, we analyze the mean Average Precision (see Table 1). Our full model noticeably outperforms the appearance-only baseline in all categories. In fact, the object-graph descriptor alone (with no appearance information) performs almost as well as our full model. For bicycles, the affinities obtained using only appearance information are weak, and thus the full model actually performs slightly worse than the object-graph descriptor in isolation. Our model's largest improvement occurs for the cow class (high appearance variance), whereas it is smaller for trees (low appearance variance).

**Impact of Known/Unknown Decisions:** Figure 7 (left) shows the precision-recall curve for our known-unknown decisions on the MSRC-v2. For this, we treat the known classes as positive, and the unknown classes as negative, and sort the regions by their entropy scores. The red star indicates the precision-recall value at $\frac{1}{2}\max E(s)$. With this (arbitrary) threshold, the regions considered for discovery are almost all true unknowns (and vice versa), at some expense of misclassifying unknown and known regions. Adjusting the "knob" on the threshold produces a tradeoff between the number of true unknowns considered for discovery versus the number of true knowns treated as unknowns. Learning the "optimal" threshold depends on the application, and for our problem setting, $\frac{1}{2}\max E(s)$ suffices.

How much better could we do with more reliable predictions of what is unknown? Figure 5 (e) shows the results for the MSRC-v2 if we replace our known-unknown predictions with perfect separation (note the vertical axis scale change). Again our model outperforms the appearance-only baseline. All purity rates are notably higher here compared to when the known/unknown separation is computed automatically, likely because the discovery problem has become much simpler: instead of having regions that could belong to one of 21 categories (total number of known and unknown categories), we only need to group the true unknowns. This implies that there is room for better initial classification (i.e., better label predictions and confidences), with which we can expect higher cluster purity rates.

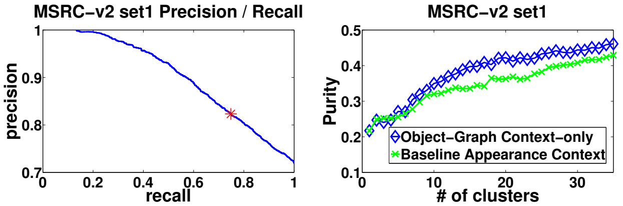

Figure 7. (**left:**) Precision-recall curve for known vs. unknown decisions on the MSRC-v2 set1; the star denotes the cutoff (half of the maximum possible entropy value). (**right:**) Comparison of the Object-Graph descriptor to a "raw" appearance-based context descriptor.

**Comparing Splits:** Upon examining the relative performance on different known/unknown splits, we found that discovery performance depends to a limited extent on which categories are known, and how many. For example, both our method and the baseline have stronger discovery performance on MSRC-v2 set2 than on set1 (see plots for set2 in supplementary file). This can be attributed to the fact that the unknowns in set2 are *grass, sky, water, road,* and *dog*, which have strong appearance features and can be discovered reliably without much contextual information. When the ratio between the number of unknown categories to known categories increases (from left to right in Figure 5 (a) and (b)), there is a decrease in the information provided by the known object-level context, and consequently we find that our improvements over the baseline eventually have a smaller margin (see rightmost curves in (a) and (b), where only 5 or 6 objects are known). Overall, however, we find that the improvements are quite stable: across the 12 random splits tested for the MSRC and PASCAL, our method never detracts from the accuracy of the baseline.

**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. We substitute class probability counts in the object-graph with raw feature histogram counts. Figure 7 (right) shows the result on the MSRC-v2. Our object-graph performs noticeably better than the baseline, confirming that directly modeling class-interactions instead of surrounding appearance cues can improve discovery.

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 appearance baseline requires $R \cdot 2Q$-dimensional vectors, for $Q$ texton + color + pHOG bins. In this case, our object-graph is $\sim$ 70 times more compact.

**Qualitative Examples of Discovered Objects:** Figure 6 shows examples of discovered categories from the 3,457 MSRC-v0 images using our approach, for $k = 30$. The cluster images are sorted by their degree as computed by the affinity matrix: $D(s_m) = \sum_{l \in L} K(s_m, s_l)$, where $L$

[3] To ensure that the improvement over [23] on the MSRC-v2 is not a result of stronger appearance features, we repeated the experiment using the same features for all methods, letting $a(s)$ be a SIFT bag of words as in [23]; our method again outperforms the baseline (see supp. file).

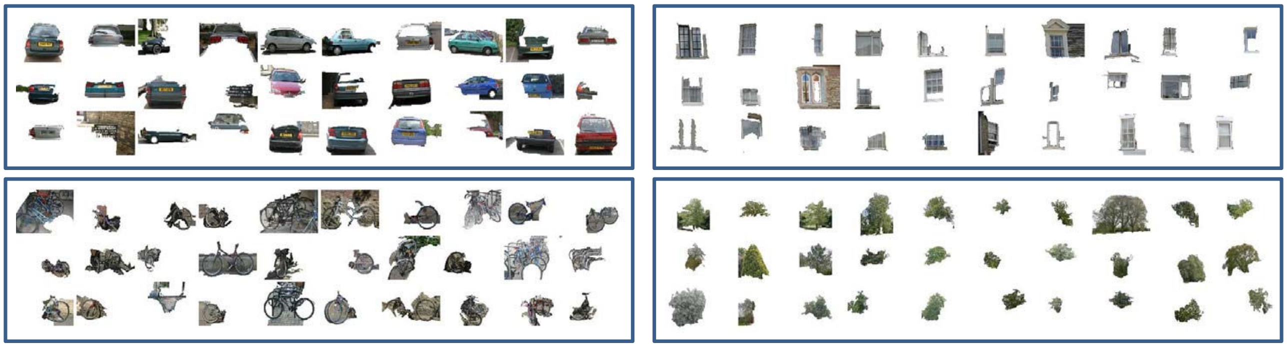

Figure 6. Examples of discovered categories for the MSRC-v0. See text for details. (Best viewed on pdf.)

denotes the cluster containing segment $s_m$. We show the top 30 regions for each cluster, removing overlapping regions and limiting to only one region per image. The resulting groups show good semantic consistency (here, we see windows, cars, bicycles and trees). Notably, our clusters tend to be more inclusive of intra-class appearance variation than those that could be found with appearance alone. For example, note the presence of both side views and rear views in the car cluster (top left), and the distinct types of windows that get grouped together (top right).

**Conclusions:** We developed an algorithm that models the interaction between familiar categories and unknown regions to discover novel categories in unlabeled images. We would like to extend the system to be used in a semi-automatic loop, where an annotator labels the meaningful discovered clusters, which would then become the familiar objects for training a classifier. We plan to next investigate ways of providing more robust known/unknown decisions.

**Acknowledgements:** We thank Alex Sorokin for helping us with the MTurk data collection, and Bryan Russell for sharing his code. This research is supported in part by NSF CAREER 0747356, NSF EIA-0303609, Texas HECB 003658-01-40-2007, the Luce Foundation, and Google and Microsoft Research.

## References

[1] F. Bach, G. Lanckriet, and M. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *ICML*, 2004.

[2] E. Bart and S. Ullman. Cross-Generalization: Learning Novel Classes from a Single Example by Feature Replacement. In *CVPR*, 2005.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. 2007.

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

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

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

[7] S. Gould, R. Fulton, and D. Koller. Decomposing a Scene into Geometric and Semantically Consistent Regions. In *ICCV*, 2009.

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

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

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

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

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

[13] H. Kashima, K. Tsuda, and A. Inokuchi. Kernels on Graphs. *Kernels and Bioinformatics*, 2004.

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

[15] S. Lazebnik and M. Raginsky. An Empirical Bayes Approach to Contextual Region Classification. In *CVPR*, 2009.

[16] Y. J. Lee and K. Grauman. Foreground Focus: Unsupervised Learning From Partially Matching Images. *IJCV*, 85, 2009.

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

[18] T. Malisiewicz and A. Efros. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. In *NIPS*, 2009.

[19] T. Malisiewicz and A. A. Efros. Improving Spatial Support for Objects via Multiple Segmentations. In *BMVC*, 2007.

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

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

[22] J. Platt. *Advances in Large Margin Classifiers*, chapter Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. MIT Press, 1999.

[23] 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.

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

[25] 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.

[26] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. 2005.

[27] A. Torralba. Contextual Priming for Object Detection. *IJCV*, 2003.

[28] Z. Tu. Auto-context and Its Application to High-level Vision Tasks. In *CVPR*, 2008.

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

[30] R. Weischedel. Adaptive Natural Language Processing. In *ACL*, 1990.