From Subcategories to Visual Composites: A Multi-Level Framework for Object Detection

Tian Lan[1], Michalis Raptis[2], Leonid Sigal[2] and Greg Mori[1]

[1]Simon Fraser University, Canada
[2]Disney Research Pittsburgh, USA

Abstract

*The appearance of an object changes profoundly with pose, camera view and interactions of the object with other objects in the scene. This makes it challenging to learn detectors based on an object-level label (*e.g.*, "car"). We postulate that having a richer set of labelings (at different levels of granularity) for an object, including finer-grained subcategories, consistent in appearance and view, and higher-order composites – contextual groupings of objects consistent in their spatial layout and appearance, can significantly alleviate these problems. However, obtaining such a rich set of annotations, including annotation of an exponentially growing set of object groupings, is simply not feasible.*

We propose a weakly-supervised framework for object detection where we discover subcategories and the composites automatically with only traditional object-level category labels as input. To this end, we first propose an exemplar-SVM-based clustering approach, with latent SVM refinement, that discovers a variable length set of discriminative subcategories for each object class. We then develop a structured model for object detection that captures interactions among object subcategories and automatically discovers semantically meaningful and discriminatively relevant visual composites. We show that this model produces state-of-the-art performance on UIUC phrase object detection benchmark.

1. Introduction

Consider the image shown in Fig. 1, humans can provide a rich set of semantic labelings for objects in such an image, including the basic object-level categories, *e.g.*, "person", the fine-grained object sub-categories, *e.g.*, "rider" and the visual contextual composite labels, *e.g.*, "person riding bicycle". Such labelings thoroughly explain the appearance variations of an object through not only the low-level pose and viewpoint changes, but also its high-level relations to other objects in this particular scene.

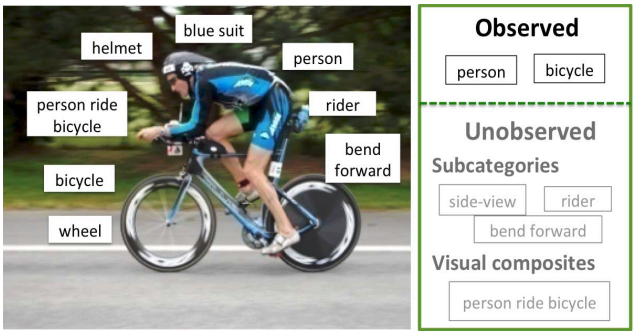

Figure 1. **Labeling an Object:** Humans can describe objects with a rich set of semantic labelings. However, in most object detection benchmarks, only the basic-level category labels are provided, *e.g.*, person, bicycle. This could result in large intra-class variations in object recognition. In this paper, we detect basic-level objects from images, and simultaneously discover the unobserved labeling space that includes the low-level fine-grained subcategory labels and the high-level visual composite labels.

Traditional object detectors perpetually struggle with a question of how to label an object. First, it is not possible to annotate (or even enumerate) every sub-category and/or composite relationship for an object category; this makes supervised training for such entities difficult. Further, such annotations are also often subjective and task specific. Second, there is a clear gap between the semantic descriptions and the discriminability of the labels for purpose of detection and classification. For example, human subjects tend to use the word "blue" to describe the person in Fig. 1, but this semantic label is not informative in classifying the instance as a "person" from other possible object categories.

To avoid burdens of annotation and issues of subjectivity, in most of the standard object detection benchmarks, only the basic-level object category labels are provided, such as

This work was supported by NSERC.

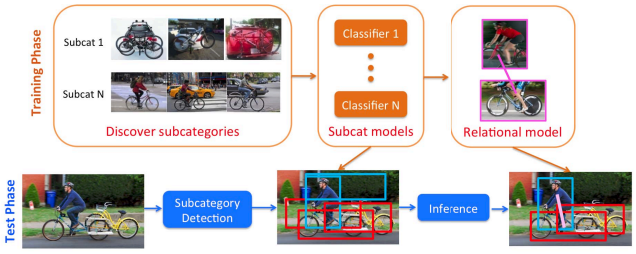

Figure 2. **An overview of our detection pipeline:** In the training phase (top row), we first discover the object subcategories from each basic-level category, then we train subcategory classifiers in the latent SVM framework. The relational model takes the outputs of these subcategory classifiers as input, and carries out reasoning on top of these responses from a structured perspective. In the test phase (bottom row), we run the subcategory detectors to generate a set of bounding boxes. Then we use our learned relational model to re-score these bounding boxes based on the object compositional relations and generate the final detection results. Our model outputs both the bounding boxes of each object and the discriminative object relations.

"person", "bicycle", "car" and "road". However, without more detailed labelings, basic-level object detectors usually suffer from large intra-class variability. Object appearance tends to change profoundly with pose, viewpoint variations and object-object interactions. For example, persons look quite different walking vs. sitting, from frontal view or side view, and their appearances also change dramatically when interacting with other objects (*e.g.*, person riding a horse or lying on a sofa has very different appearance from a generic person walking). We believe the multi-level (contextual) semantic labelings are key to resolving intra-class variations within the basic-level object categories. Considering the difficulties in obtain such labelings, we advocate a weakly supervised setting where only the basic-level categories are provided in training, and the fine-grained *subcategories* as well as the high-level *visual composites* are automatically discovered from the training data.

In this paper, we propose a novel framework for detecting and labeling objects with basic object-level categories and multiple automatically discovered semantic labelings. The semantic labelings consist of lower-level object subcategories as well as higher-level visual contextual composites, modeling relationships that the object has with other objects in the scene. Object subcategories are visual clusters that capture a wide range of appearance variations of an object. We propose an exemplar-SVM based clustering algorithm to discover the subcategory labels. The subcategories are then treated as mixture components in a latent SVM framework, and refined during learning. Our model detects possibly multiple object instances in a single image and generates detailed fine-grained subcategory labels for each instance. At the higher-level of our framework, we focus on the reasoning about relationships between object subcategories. Such reasoning generates sets of closely interacting objects that form visual composites, *e.g.*, "person riding horse". Notably the structure of the composites, including participating objects and their spatial layout, are discovered automatically from data using a structured model formulation. An overview of our approach is shown in Fig. 2.

	Multi-component Object Detection		
	Model	**Clustering**	**# of comp**
Bourdev and Malik [1]	SVM	annotations	fixed (300)
Divvala *et al.* [16]	(part-based) LSVM	appearance	fixed (15)
Divvala *et al.* [17]	(part-based) LSVM	appearance	fixed (25)
Felzenszwalb *et al.* [7]	part-based LSVM	aspect	fixed (3)
Gu and Ren [11]	part-based LSVM	view	fixed (4/8)
Gu et al. [10]	SVM	annotations	400-1000
Malisiewicz *et al.* [14]	SVM	none	all exemplars
Our model	part-based LSVM	**appearance**	**data-driven**
	automatic	**spatial subcategory co-occurrence**	
Li *et al.* [13]	automatic	spatial object co-occurrence	
Sadeghi *et al.* [15]	none	manual annotation	
	Discovery	**Patterns**	
	Visual Composites		

Table 1. **Related Work:** Our framework relates to both multi-component object detection and, recently introduced, contextual visual composite models. A notable difference is that we *discover* sub-categories and composite relationships automatically from data (based on appearance and spatial layout).

Contributions: 1) We propose a spectrum of automatically generated semantic labelings for object detections. These labelings contain both lower-level subcategories and higher-level visual composites. 2) We introduce a discriminative clustering algorithm to discover the subcategories. 3) We develop a structured model for object detection that captures interactions among object subcategories and automatically discovers discriminative visual composites. We also show that our discovered visual composites are semantically meaningful. 4) Our approach produces state-of-the-art performance in an object detection benchmark that includes a rich set of object interactions.

2. Related Work

Our work relates to a number of topics in object detection and recognition. We overview the closest literature in Table 1 and discuss it further below.

Multi-component Object Detection: In order to deal with significant intra-class variations that can not be tackled by monolithic models, several influential approaches modeling multiple components of objects, *i.e.* subcategories, have been introduced [1, 7, 10, 11, 14, 16]. Mixture components were integrated into the deformable part models either based on bounding box or appearance k-means clustering in [7] and [16, 17] respectively. However, the number of mixture components is pre-defined (fixed) and not inferred from data. Gu and Ren [11] focus on modeling viewpoint variations of objects and ignore other richer sources of intra-class variations. Malisiewicz et al. [14] train an exemplar SVM for each positive example. However, the generaliza-

tion ability of each model is limited. In this work, we use exemplar SVM to discover an initial pool of subcategories which we then refine, through merging, and train part-based models for each resulting subcategory. Gu et al. [10] is similar to our subcategory model. However, in [10] object relationships are not modeled and hand annotated keypoints and masks are used for clusterings, making the approach considerably less scalable. Bourdev [1] introduces poselets for person detection, but poselets encode visual composites of parts rather than global objects and also require keypoint annotation.

Object Interactions: There is rich literature on modeling contextual interactions between objects, including [2, 4, 9, 12, 19]. The core difference between our approach and these methods is that we model interactions among object *subcategories* rather than the basic-level object categories. This captures the subtle joint appearance changes between objects caused by interactions. For example, when a person partakes in an interaction with a bicycle, when riding, the appearance of both the bicycle and the person exhibit view-consistent appearance changes (*e.g.*, the rider's legs tend to occlude specific parts of the bicycle and the bicycle creates a highly textured background close to the rider's legs).

Visual Composites: Recently, several works implicitly model occlusions and interactions through entities that fall between objects and scenes. This is often referred to as "visual composites" – two or more closely interacting objects. Sadeghi *et al*. [15] manually annotate a list of visual phrases and train global phrase templates for detection. In [13], higher-order visual composites are automatically discovered based on the spatial/scale/aspect consistency of objects. However, the appearance consistency of composite visual patterns are not taken into account. These global template based approaches may require a separate template for each combination of interacted objects, making scalability problematic. In contrast, we use subcategories and spatial relations to reason about object interactions. Desai and Ramanan [3] propose phraselets, where human pose is modeled together with interacting objects based on the configurations of local patches. Notably, our work models interactions among object subcategories.

3. Discovering Subcategories

Given a set of training images with basic-level object category labels and bounding boxes, our goal is to discover the fine-grained subcategories. The two key requirements for good object subcategories are: (1) inclusivity – subcategories should cover all, or most, variations in object appearance and (2) discriminability – subcategories should be useful for detecting the class. The standard solution is to employ some form of unsupervised clustering, such as k-means on the object appearance feature vectors [16, 17]. However, running k-means on objects usually does not produce good clusters, particularly in terms of discriminability, due to the low-level predefined (Euclidean) distance metric used by k-means. In addition, manually defining the number of clusters is often difficult. We argue that the number of subcategories per object class should be driven by the appearance variations within that class, not a fixed global parameter. An alternative option is the unsupervised mid-level patch discovery strategy proposed recently in [18]. The method shows good performance in finding a set of representative patches from unlabeled images.

Our approach is inspired by the recent success of exemplar SVM [14]. We train a linear SVM for each exemplar. The exemplar is used as the single positive example, while negative examples are sampled from images that do not contain any instances of the exemplar's class. We use 5 iterations of hard negative mining in training each classifier. An exemplar is represented by a rigid HOG template, and each classifier can be interpreted as a learned exemplar-specific template. For each exemplar, we run the detector on all other examples in the class. We consider detection scores above -1 indicative of object presence. A cluster is formed by the exemplar and its top k scoring detections. We limit each cluster to only five members ($k = 4$) to keep cluster homogeneity. Clusters with fewer than two members are pruned. This process allows us to obtain a large set of highly homogenous atomic clusters. We then merge visually consistent clusters via affinity propagation [8]. We define the (asymmetric) similarity from cluster s to cluster r as: $d_s(r) = \frac{1}{\mathcal{N}} \sum_i \sum_j w_s(i)^\top \mathbf{x_r}(\mathbf{j})$, where $\mathbf{x_r}(\mathbf{j})$ is the HOG feature vector of the j-th example of the cluster r. The weight term $w_s(i)$ is the learned template for the i-th example of the cluster s. $\mathcal{N}$ is the normalization constant computed as the number of examples in cluster r times the number of examples in cluster s.

We compute similarities between every pair of the atomic clusters. Then affinity propagation [8] is applied where the atomic clusters are gradually merged into larger clusters by a message-passing procedure. Unlike k-means, affinity propagation does not require parameters that specify a desired number of clusters, instead the number of clusters is determined from the data. Affinity propagation also does not require initialization of cluster centers. We run affinity propagation for each basic-level category separately and obtain the fine-grained subcategories. Figure 3 shows a visualization of several example subcategories. Note that, due to our discriminative training strategy, objects within each subcategory are highly consistent in appearance.

4. Learning Subcategories

Given the set of subcategories obtained from the previous section, we learn a mixture model based on DPM [7], where the mixture components correspond to subcategories

Figure 3. **Subcategories:** Subcategories are defined by object instances that are tightly clustered in appearance space. The figure shows the discovered subcategories for some of the basic-level object categories including horse, person, bicycle, car, bottle and dog. We only show the first five examples in each subcategory.

of a basic-level object category. We first review the key methodologies of learning the DPM detector, and then explain the details of their use in our mixture model.

DPM is trained from a set of labeled examples $D = (\langle x_1, y_1 \rangle, \ldots, \langle x_n, y_n \rangle)$, where $y_i \in \{-1, 1\}$. The goal is to learn model parameters w by minimizing the objective function,

$$L_D(w) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i f_w(x_i)) \quad (1)$$

where $f_w(x_i)$ is the score of the classifier w on example x_i. Here we write $f_w(x) = \max_{z \in Z(x)} w^\top \Phi(x, z)$; z are the latent variables and $Z(x)$ are the set of possible latent values for an example x. Since the optimization is non-convex, the model parameters w and the latent variables z are learned using an iterative approach [7].

Mixture models: It is straightforward to train DPM for each subcategory independently [16]. However, one concern is in calibrating the scores output by individual SVM classifiers. In addition, subcategories discovered in the previous section might be noisy and should be cleaned up during learning. In this work, we train the subcategory classifiers in the latent SVM framework, where the training of the classifiers are coupled and the subcategory labels are refined in the latent step. The subcategory labels in our method correspond to the mixture components in DPM.

In DPM, the mixture components are initialized according to the examples' aspect ratios and updated in the latent step. However, the aspect ratio heuristic does not generalize well to a large number of subcategories, and thus often fails to provide a good initialization. Due to the non-convex nature of latent SVM, initialization of subcategories is a key step of learning a good detector. Here we naturally use the subcategories discovered in the last section to initialize the mixture components, and allow the subcategory labels to refine during the latent step. To detect objects, we run the learned classifiers independently for each component. The output of this step is a set of candidate windows where each window is associated with a fine-grained subcategory label.

5. Relational Model

Subcategory mixture models learned in the previous section are good for dealing with appearance variations present in a given object class, however, detection accuracy for some, typically smaller and less-discriminative, object classes may still be low. Contextual information of relationships between detections in an image can further boost the performance of object detectors, as been shown in [13] and [15]. Unlike these methods, however, we propose to build contextual basic-level category models based on the subcategory classifiers (not aggregate object detections, as in [13], or composed templates [15]). This allows our model to be attuned to visual and view-based correlations between subcategories of objects.

We use a star model to represent an object, where the object is connected to other objects in the vicinity of the detection. Intuitively, this is similar to the part based model in [7], and the objects in context are treated as parts. However, instead of treating all objects in the same image as context, we introduce binary latent variables to discriminatively select which objects have strong interactions with the central object and should be included in our model. We call the star graph that includes the candidate central object and the contextual objects a *visual composite*. At the end of the inference, our goal is to obtain a rich set of visual composites that are not only highly characteristic of the object class, but also highly discriminative compared to other classes.

Representation: We begin by introducing the notations we use in the rest of the paper. The input to our learning module is N images accompanied with a set of $\langle X_n, Y_n \rangle$ pairs, $n = 1, \ldots, N$. Here we write $X_n = \{x_i : i = 1, \ldots, M_n\}$ as the representation of the n-th image, where M_n is the total number of detected bounding boxes for this image and x_i is the feature vector of the i-th bounding box. Let $Y_n = \{y_{c,i} : c = 1, \ldots, C, \text{and } i = 1, \ldots, M_n\}$ be the entire label set for image X_n, where C is the total number of object categories in the dataset and $y_{c,i} \in \{0, 1\}$ is a variable indicating if the i-th bounding box contains an object of the c-th category. Let $p_i \in \{1, \ldots, K\}$ be the indicator variable showing the subcategory detector that select the i-th bounding box, where K is the total number of subcategories in our dataset. An object hypothesis is represented by a star graph, which specifies an object bounding box and a set of bounding boxes of contextual objects. Assuming the central object's index is i, we use L_i to represent the indices of contextual objects in an image where $L_i = \{j : j \in \{1, \ldots, M_n\} \backslash i\}$.

5.1. Finding Candidate Visual Composites

In training, we discover a set of discriminative visual composites automatically. We define visual composites as consisting of two or more objects. Objects that belong to the same composite should co-occur frequently and conform to certain spatial and scale relationships that are consistent across images. Based on this definition, only a subset of objects in an image can be part of a visual composite. Our goal is to discover such composites that exhibit consistently occurring object layout patterns in a set of images.

Consider a visual composite, represented by a star graph. The key insight is that based on our definition, contextual objects windows (leaves of the graph) should all have consistent layout with the central object window; in other words, contextual objects should be able to consistently predict a bounding box for the central object under consideration. With that as an insight, at training time, we first learn spatial layout relationships between potential contextual objects and central object. Based on these learned relations an initial graph of a visual composite is constructed (by only considering object detections that are consistent in predicting considered central object's bounding box).

We start by fitting a three component Gaussian mixture (MoG) model to pairs of (bounding boxes of) objects that co-occur in training images. The three component MoG allows us to model various spatial and scale aspects of the object-object relationship. Notably, we can easily produce a hypothesis for a bounding box of a central object by conditioning the learned mixture model on the bounding box of a contextual object. Given an image, we can use this model to determine the set of possible contextual objects for each central object window. Given a central object window we consider contextual objects to be windows that, given a learned spatial Gaussian mixture model, can predict the central object window to > 0.3 overlap (VOC criterion).

During *training*, we iterate over all true positive activations (responses of the detector that are within 0.5 overlap to the true object annotation), and in this way obtain the visual composites (central object + contextual objects) that have tight spatial configuration coupling.

During *testing*, however, the ground truth object categories are not provided, thus we do not know if a detection window is a true positive or not. If we naively include all spatially consistent detection windows as contextual objects for a given candidate central object, we may include many false positives and thus hurt the performance. Thus during inference, we introduce a binary latent variable for each candidate contextual window to discriminatively select if it will be included in our composite object model. For an object window i, we use $\mathbf{h}_i$ to denote the binary latent variables for all contextual objects with indices in the set L_i. Note, during training we assume $\mathbf{h}_i$ is known (see above).

5.2. Model Formulation

We construct models for each basic-level object category separately. For modeling the c-th basic-level object category, the score associated with a bounding box i is:

$$S_c(x_i, y_{c,i}, \mathbf{h}_i) = \alpha_{p_i}^\top x_i \cdot y_{c,i} + \sum_{j \in L_i} \beta_{p_j}^\top d_{ij} \cdot h_{ij} + \sum_{j \in L_i} \gamma_{p_i p_j}^\top x_j \cdot h_{ij} \tag{2}$$

Root model $\alpha_{p_i}^\top x_i \cdot y_{c,i}$: We simply use the output of the subcategory detector as the single feature. To learn biases between different subcategories, we append a constant 1 to make x_i two-dimensional. α_{p_i} is the two-dimensional weight that corresponds to the subcategory class of the i-th bounding box. If the bounding box is labeled as background $(y_{c,i} = 0)$, then the potential of the root model is set to zero.

Context model $\beta_{p_j}^\top d_{ij} \cdot h_{ij}$: We write $d_{ij} = [x_j,\ g_{ij}]$ to represent the objects in context, where x_j is the appearance feature (detection score) of the j-th bounding box and g_{ij} is the spatial feature computed based on the relative position and scale of the j-th bounding box w.r.t. the i-th bounding box using the Gaussian distribution described in the previous section. h_{ij} is a binary latent variable that determines whether the contextual object is discriminative and should be included into the context model.

Co-occurrence model $\gamma_{p_i p_j}^\top x_j \cdot h_{ij}$: This term captures the "prior" over subcategory combinations. The intuition is that certain pairs of subcategories tend to co-occur while others do not, for example, a bicycle with side view tends to co-occur with a rider with the same viewpoint, and a horse tends to co-occur with a horse rider instead of a person walking.

5.3. Inference

We assume that the bounding box labels $y_{c,i}$ are independently inferred, and our inference is exact. For an object window i, our inference corresponds to solving the following optimization problem:

$$(\widehat{y_{c,i}}, \widehat{\mathbf{h}_i}) = \arg\max_{y_{c,i}, \mathbf{h}_i} S_c(x_i, y_{c,i}, \mathbf{h}_i) \tag{3}$$

For the bounding box i, the inference is on a star graph where we jointly infer the presence or absence of the c-th category $y_{c,i}$ as well as the corresponding binary latent variables $\mathbf{h}_i$ of the contextual objects. This is very simple exact inference as $y_{c,i}$ and $\mathbf{h}_i$ are all binary variables and we can enumerate all possible values of the random variables and find the optimal solution. Note that we constrain $\mathbf{h}_i$ to an all-zero vector when $y_{c,i} = 0$, which means we do not consider object interactions with background. We emphasize that our inference procedure returns both object labels and the visual composites that tells the closely related contextual objects for each object window.

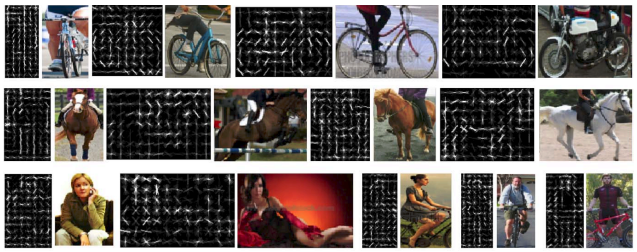

Figure 4. **Subcategory Templates:** Examples of learned three object categories: bicycle, horse and person (one category per row). For each template, we show the best match from the UIUC phrase training set. For object categories, bicycle, bottle, car, chair, dog, horse, person and sofa, the number of discovered subcategories (through affinity propagation) for our mixture models are: 13, 6, 15, 6, 13, 12, 59 and 13 respectively. Due to space limitation, we only show some of the templates.

5.4. Structure Learning

Given N training images with a set of bounding boxes X and the corresponding object category indicator labels Y. We would like to train the model parameter θ that tends to produce the correct object labels. Here we train models independently for each basic-level object category. The following objective function is for learning the model parameter θ_c for the c-th category:

$$\min_{\theta_c, \xi \geq 0} \frac{1}{2}||\theta_c||^2 + C \sum_{n=1}^{N} \sum_{i=1}^{M_n} \xi_i^n$$
$$S_c(x_i^n, y_{c,i}^n, \mathbf{h}_i^n) - S_c(x_i^n, \widehat{y_{c,i}^n}, \widehat{\mathbf{h}_i^n}) \geq \Delta\left(y_{c,i}^n, \widehat{y_{c,i}^n}\right) - \xi_i^n, \quad \forall i, \forall n, \quad (4)$$

where the loss function Δ is a 0-1 loss that measures the difference between the ground-truth object category indication $y_{c,i}$ and the inferred variable $\widehat{y_{c,i}^n}$, *i.e.*, $\Delta\left(y_{c,i}^n, \widehat{y_{c,i}^n}\right) = 1$ if $\widehat{y_{c,i}^n} \neq y_{c,i}^n$, and 0 otherwise. This form of learning problem is known as structural SVM, and many well-tuned solvers can be applied to solve this problem. Here we use the bundle optimization solver in [5].

We note that during training the contextual binary variables $\mathbf{h}_i$ are observable based on the ground truth visual composites (see Section 5.1). We experimented with letting $\mathbf{h}_i$ be latent during learning, however, saw somewhat inferior performance which we attribute to resulting non-convexity in the objective.

6. Experiments

We present results of object detection on a standard object benchmark dataset: UIUC phrase dataset [15]. The UIUC phrase dataset contains 2796 images which consist of a subset of PASCAL images and images for phrases collected from the web. The images are labeled with 8 of the 20 PASCAL categories, and a list of 17 visual phrases such as person riding bicycle, dog lying on sofa, *etc*. In this paper, we train our object models using only the basic-level 8 category labels. We use the training-testing split avaialble at **h**ttp://vision.cs.uiuc.edu/phrasal/.

We picked UIUC phrase because it contains a rich set of visual composites (phrases plus a few higher-order composites such as person drinking bottle sitting in a chair). It also contains a subset of PASCAL images and uses familiar PASCAL object categories, PASCAL on the other hand contains few composites. The use of UIUC phrase dataset also allows us to compare our results to competing methods. We compare our results to state-of-the-art performance results of [13, 15], as well as the detection performance of our subcategory classifiers. We further apply the proposed method to image retrieval with visual phrase queries, and show that our method significantly outperforms baselines.

6.1. Object Detection

Table 2 compares results of our method with leading approaches on the UIUC phrase dataset. The approaches we compared against are: 1) Deformable part-based model (DPM) [7]; 2) Object context: a contextual re-scoring scheme used in [7]; 3) Phrase context: object detection outputs are re-scored using the phrase template trained by manually defined phrases [15]; 4) Group context [13], object detection outputs are re-scored using automatically discovered groups of objects. To fairly compare with the reported results, we use the same version of deformable part models to train our subcategory classifiers [6].

Our method achieves state-of-the-art in terms of mean average precision across 8 object categories on this benchmark. In particular, for 4 object categories: car, dog, horse and person, our method significantly improves on state-of-the-art [7, 13] by $5.5\%, 9.4\%, 11.2\%$ and 12.1% respectively. Note that our method does not use phrase annotations as in [15], but achieves significantly better performance. There are two main reasons for the improved performance in our method: 1) We discover highly consistent subcategories. The trained subcategory classifiers are highly discriminative and address intra-class variations among basic-level categories. This is demonstrated by the performance gain of subcategory classifiers in Table 2. Some of the learned subcategory templates are visualized in Figure 4. 2) Modeling visual composites improve performance for some hard-to-detect object categories such as bottle and chair. We think the poor performance of classifying bottle and chair by our subcategory classifiers is mainly due to over-fitting. We have fewer training examples for these two classes, and further dividing these examples into subcategories will easily over-fit the data. In this case, the context of visual composites is important. For example, a person with a pose of sitting will help detect a chair below him. The visual com-

	bike	bottle	car	chair	dog	horse	person	sofa	mAP
DPM [7]	57.0	7.0	25.8	11.1	5.6	49.3	25.7	14.1	24.5
Object context [7]	58.8	9.3	33.1	13.4	5.0	53.7	27.9	19.8	27.6
Phrase context [15]	60.0	9.3	32.6	**13.6**	8.0	53.5	28.8	22.5	28.5
Group context [13]	63.5	**10.7**	32.5	13.2	8.0	54.6	30.6	**24.9**	29.8
Our subcategory model	63.7	3.0	37.7	4.7	14.1	**67.0**	**45.2**	23.7	32.4
Our full model	**63.9**	9.4	**38.1**	9.8	**17.4**	65.8	42.7	24.4	**33.9**

Table 2. **Detection results on UIUC phrase:** Table compares average precisions (AP) for all 8 categories and the mean AP (mAP) across categories; leading approach is in **bold**. Our method achieves state-of-the-art in mean AP and outperforms [13, 15] in 5 out of 8 categories.

posites will also prune some false positives that violate typical spatial configurations. A detailed discussion of visual composites is provided in the next section.

6.2. Visual Composites

Fig. 5 visualizes our detection results along with the predicted visual composites. For example, in Fig. 5 (a), a confident "person" response together with a tight spatial layout helps boost the bicycle detection. Compared to phrases, higher order composites can be more effective because multiple confident contextual object responses in tight spatial layout with the single central object detection tend to be more reliable and indicative. Fig. 6 (top right) shows a composite of person riding bicycle with a car nearby, where the confident responses of car and person will jointly boost the detection of the bicycle.

Visual phrase retrieval: Since the UIUC phrase dataset is annotated with visual phrases, we wish to evaluate how well our method performs in applications designed for those phrases even without phrase annotations during training. However, our method is not directly applicable to visual phrase detection as in [15], since the output of our method are bounding boxes of objects instead of visual phrases. Here we evaluate our method in image retrieval with visual phrase queries. Instead of answering question of where is the visual phrase, we focus on whether an image contains the visual phrase.

UIUC phrase contains annotations of 12 visual phrases that describe interactions between two objects (*e.g.*, person lying on sofa). We evaluate performance by using each visual phrase as the query. We compare our method with deformable part models [7] and our subcategory detectors. All of the three methods are trained with the same annotations – only the basic-level object categories are provided. Since these methods cannot make a distinction between phrases composed of the same objects (*e.g.*, "person lying on sofa" versus "person sitting on sofa"), while evaluating the retrieval performance for a query (*e.g.*, "person lying on sofa"), we remove all images from the test set which contain other queries that are composed of the same objects (*e.g.*, "person sitting on sofa").

We use the same heuristic to combine the object detector scores for retrieval. For example, given a query of "person riding bicycle", we first apply both "person" detector and "bicycle" detector on the testing images. Then a testing image will be scored by the sum of the maximum person detector score and maximum bicycle detector score on this image. Note, detector scores across object categories have been normalized to the same scale by logistic regression.

Phrase Names	DPM [7]	Subcat.	Full Model
Person lying on sofa	1.0	**2.8**	**2.8**
Person sitting on sofa	4.6	9.0	**12.4**
Person riding bicycle	67.4	86.5	**86.9**
Person next to bicycle	45.9	59.3	**68.0**
Person riding horse	78.3	**85.8**	85.7
Horse and rider jumping	15.7	70.5	**71.3**
Person next to horse	27.9	**28.2**	27.4
Person drinking bottle	**8.4**	2.5	2.1
Person sitting on chair	**11.0**	10.3	10.7
Person next to car	18.7	**36.1**	35.6
Bicycle next to car	20.7	**56.7**	54.2
Dog lying on sofa	4.6	18.4	**20.4**
mAP	24.3	38.9	**39.8**

Table 3. **Phrase query based image retrieval results:** Comparison of average precisions (AP) for all 12 visual phrases and the mean AP (mAP) across categories; leading approach is shown in **bold**. Our method significantly outperforms the baseline and the full model improves over our subcategory model.

Summary of results is in Table 3. Our method again achieves remarkable improvement over the baseline. Note that our method does not require any visual phrase annotations, and is trained purely on basic-level object categories, but can still reliably retrieve a list of visual phrases: "Person riding bicycle", "Person riding horse" and "Horse and rider jumping". The biggest performance gap is on "Horse and rider jumping": our method increases the average precision from 15.7% (DPM) to 70.5%. We believe this is because our method automatically discovers subcategories corresponding to "horse rider" and "jumping horse" and learns discriminative subcategory templates. Our full model also further improves the results of subcategory classifiers, particularly in the phrases with tight spatial configurations, such as "Person sitting on sofa" and "Horse and rider jumping". Our method does not perform well for "person drinking bottle". We believe this is because our subcategory classifiers for bottle over-fit (see Section 6.1), and the response of bottle detector during testing is often low; notably, it can

Figure 5. **Object and visual composite detection:** The central objects, object we want to re-score, are shown in red (only the top 3 detection responses are visualized), and the automatically discovered contextual objects in blue; green rectangles label the visual composites. For each detection response, we also show the confidence score before and after applying our relational model denoted by $s : t$, where s is the output of subcategory detector and t is the output of our full model. We use dashed line to denote the responses suppressed by our relational model, solid line for boosted responses. For example, in (a), a confident "person" response together with tight spatial layout boosts the "bicycle" score from 0.99 to 1.31 while the false positive bicycle response above the person is suppressed; (b) shows an example of "person-bottle" composite. Although the true positive bottle response is decreased, the gap between true positive and false positive bottles is increased.

easily be dominated by the high response of person since the confidence score of "person drinking bottle" is obtained by summing "max" scores of person and bottle.

7. Conclusion

In this paper, we propose a multi-level framework for detecting and labeling objects with basic object-level categories and multiple automatically discovered semantic labelings including the fine-grained subcategories as well as the high-level visual composites. Our framework is weakly supervised where only the basic-level categories are provided in training. Our experiments on the UIUC phrase dataset show that the proposed method outperforms multiple state-of-the-art methods in object detection, and the automatically discovered visual composites are semantically meaningful. We further show that our method can be applied to retrieve images with visual phrase queries even without visual phrase annotations during training, with significant improvement over baselines.

Figure 6. **Visual composite examples:** The central objects are in red, contextual objects are in blue. For clarity, we omit to show relative spatial configuration. As can be seen, composites are often are semantically meaningful. Besides visual phrases, we also discover higher-order composites such as: person-bicycle-car and person-bicycle-person (see examples of the second row).

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors training using 3d human pose annotations. In *ICCV*, 2009.

[2] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[3] C. Desai and D. Ramanan. Recognizing actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, 2009.

[5] T.-M.-T. Do and T. Artieres. Large margin training for hidden markov models with partially observed states. In *ICML*, 2009.

[6] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://people.cs.uchicago.edu/ pff/latent-release4/.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1672–1645, 2010.

[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 2007.

[9] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, 2010.

[10] C. Gu, P. Arbelaez, Y. Lin, K. Yu, and J. Malik. Multi-component models for object detection. In *ECCV*, 2012.

[11] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[13] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012.

[14] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[15] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[16] M. H. Santosh K. Divvala, Alexei A. Efros. How important are 'deformable parts' in the deformable parts model? In *Parts and Attributes Workshop, ECCV*, 2012.

[17] M. H. Santosh K. Divvala, Alexei A. Efros. Object instance sharing by enhanced bounding box correspondence. In *BMVC*, 2012.

[18] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[19] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.