Attribute-Centric Recognition for Cross-category Generalization

Ali Farhadi Ian Endres Derek Hoiem
University of Illinois at Urbana Champaign
{afarhad2,iendres2,dhoiem}@illinois.edu

Abstract

We propose an approach to find and describe objects within broad domains. We introduce a new dataset that provides annotation for sharing models of appearance and correlation across categories. We use it to learn part and category detectors. These serve as the visual basis for an integrated model of objects. We describe objects by the spatial arrangement of their attributes and the interactions between them. Using this model, our system can find animals and vehicles that it has not seen and infer attributes, such as function and pose. Our experiments demonstrate that we can more reliably locate and describe both familiar and unfamiliar objects, compared to a baseline that relies purely on basic category detectors.

1. Introduction

Researchers have made great progress in developing systems that can recognize an individual object category. But what if we want to recognize many objects? The current solution is to build a new detector for each category of interest. While simple, this approach does not acknowledge the commonalities among many different types of objects. One consequence is inefficiency: each new detector requires many training examples, and evaluation time grows linearly. But the main downside of the approach is that each category needs to be defined in advance. This is a major problem for many applications. For example, an automated vehicle needs to recognize a horse in the road as an animal and predict its movement, even if it has never seen one (Figure 1).

In this paper, we propose a more flexible and integrative framework that enables new objects to be understood with respect to known ones, allowing them to be partially recognized. Instead of learning each category separately, we group objects within broad domains, such as "animal" and "vehicle". During training, we learn detectors for categories and parts that are shared across basic-level categories. For example, "leg" and "four-legged animal" detectors are shared by dogs and horses, while the "dog" detector applies only to dogs. Using a simple graphical model, we also encode the correlations among attributes, including parts, categories, pose, and function. Through a shared representation, we enable the system to predict that a horse is a four-legged animal and that is standing and capable of walking, even if it has not seen any horses during training. During testing, our system finds new objects by voting for object locations using our part and category detectors. Using the learned correlations, the system then infers the attributes of the object and the likelihood that it is within a known domain.

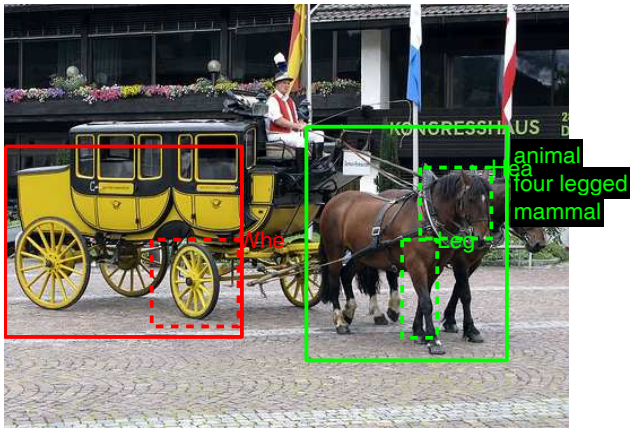


Figure 1. In this result, our system has never seen a horse or a carriage, but it is able to localize them and identify their parts, among other attributes.

Our goal is to find and describe *any* object within known domains. This ability to generalize beyond specifically trained tasks is crucial for many applications, but existing recognition datasets are designed only for study of individual category recognition. Accordingly, we provide a new CORE (Cross-category Object REcognition) dataset that allows development and study of object models with intermediate semantics. Our dataset includes 2,800 images of natural scenes, with segmentations and attribute annotations for objects in 28 categories of vehicles and animals. We train on 19 categories of *familiar objects* and test on new images containing all 28 categories, including 9 categories of *unfamiliar objects* whose basic-level categories are not seen during training.

We perform experiments on two tasks: 1) find all animals and vehicles; and 2) assign attributes to localized objects. We compare our method that integrates shared detectors and reasoning about attributes to a baseline that detects basic-level categories and infers attributes directly from the cate-

gory. Our model outperforms the baseline by a surprising margin for both tasks, improving recognition of familiar objects and doubling the recall of unfamiliar objects at a fixed false positive rate.

Background. The earliest works in object recognition attempted to model objects in terms of configurations of shared materials, parts, or geometric primitives [32, 14, 6, 15, 26, 25, 4, 31]. Ultimately, these methods gave way to simpler, more direct and data-driven methods for recognition that avoid hand-coded models. We now have several advantages that make it propitious to revisit recognition with intermediate semantics. First, researchers have made great strides in basic pattern matching. We show that an existing detector from Felzenszwalb et al. [13] can learn appearance models of parts and objects that perform well in our difficult dataset. Second, digital images are abundant, enabling data-driven, statistical approaches and rigorous evaluation. Finally, annotation is now also easy to obtain, with services such as Amazon's Mechanical Turk [34]. With an abundance of data, fast computers, large-scale annotation services, advanced machine learning methods, and improved low-level features, we believe that object representation is the key to progress in recognition.

Our focus is on creating the right level of abstraction for knowledge transfer. Others [37, 27, 35, 20, 36, 7, 2, 12, 3, 22] have shown that sharing low-level features can improve efficiency or accuracy, when few examples are available. But on challenging datasets [10] with many training examples, these methods have not yet been shown to outperform the best independently trained detectors (e.g. [13]). By providing stronger supervision, we enable more effective knowledge transfer, leading to substantially better performance than standard object detectors at localization and naming, while additionally inferring pose, composition, and function.

In our use of supervised parts to aid detection, we relate to recent works on learning compositional models of objects [40, 16, 39, 1]. Compositional models are attractive because they allow different objects to be represented by shared components, allowing learning with fewer examples. Though our aim relates, our models are much simpler, and we are able to achieve state-of-the-art results on a difficult dataset.

Our aim to improve generalization through supervised intermediate semantics is related to several recent works. Palatucci et al. [28] study the generalization properties of systems that use intermediate representations to make predictions for new categories, with application to interpretation of neural patterns. Kumar et al. [17] show that predicted facial attributes, such as fullness of lips, are highly useful in face verification. More generally, their work demonstrates the role of intermediate semantics for subcategory differentiation, while ours focuses on generalization across broad domains. Farhadi et al. [11] and Lampert et al. [18] show that supervised attributes can be transferred across object categories, allowing description and naming of objects from categories not seen during training. These attributes were learned and inferred at the image level, without localization. In contrast, we learn localized detectors of attributes and encode their spatial correlations. This allows us to automatically localize objects and to provide much more accurate and detailed descriptions.

Contributions. Overall, we demonstrate the promise of an approach that infers an underlying semantic representation through shared detectors. By learning about one set of animals or vehicles, we can localize and describe many others. This ability is essential when a system must reason about anything it encounters. In the past, limited availability of data and annotation has hindered attempts to learn more integrated models. Our dataset should make such studies much more feasible. In summary, this paper offers the following contributions:

- Framework for more flexible and integrative recognition that allows objects within broad domains to be localized and described
- Techniques for knowledge transfer of appearance, spatial and relational models
- CORE dataset that enables development and study of object models with intermediate semantics
- Validation of our approach and study of how well appearance-based detectors of parts and superordinate categories can generalize across object classes

2. Learning Shared Object Models

We have created a new dataset for studying shared representations and cross-category generalization. We use it to learn shared appearance models, co-occurrence, and spatial correlations.

2.1. Dataset

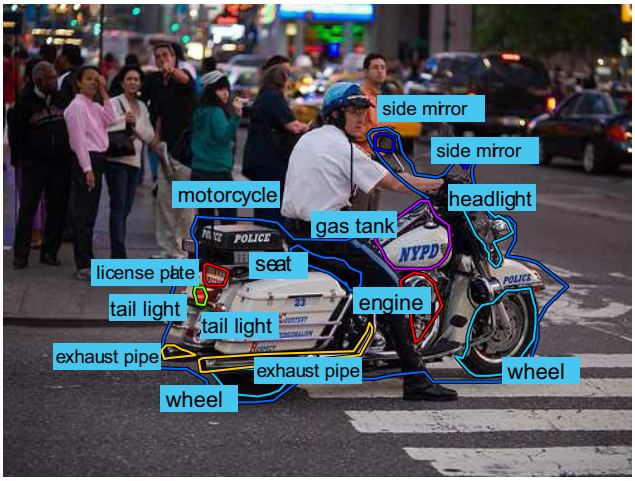


Figure 2. Example of an annotation in our dataset.

Representation is the key to effective knowledge transfer. Observations of biological systems suggest that good representations can be learned automatically, leading to much research in unsupervised discovery of latent structure in images or objects. However, for a passive machine that cannot explore or manipulate objects, it is not known whether such structure can be discovered from images without supervision.

To allow exploration of strongly supervised approaches and shared representations, we have created a new CORE (Cross-category Object REcognition) dataset. We currently have roughly 3,000 annotated objects in 2,800 images, many gathered from ImageNet [9]. The annotations for each object include object segmentation, segmentation of parts, category and part labels, masks for common materials, pose, and viewpoint. In total, 28 different kinds of objects (animals and vehicles) are annotated, as well as several dozen types of parts and ten materials. We used labelers in Mechanical Turk with careful quality checks. Our annotation is motivated, in part, by research in human concepts and categories [33, 24, 30, 23]. We show an example of an annotation in Figure 2, and typical scenes can be seen throughout in our result figures. In our work, we use only a subset of the annotation: bounding boxes, names of objects, parts, poses, and functional attributes. The dataset and annotations along with supporting code, documentation and detector models are currently available at http://vision.cs.uiuc.edu/CORE.

In comparison to PASCAL VOC [10], our dataset appears to be slightly easier for basic-level object detection (average AP using [13] is slightly higher), likely because our dataset has fewer occluded vehicles, but our dataset includes several very difficult categories (bat, whale, and boat) with AP less than 0.05. However, our dataset is intended to study the much greater challenge of cross-category generalization in localization and description.

2.2. Shared Appearance Models

Shared appearance models are the foundation of our approach. If we cannot detect parts or objects, even the most sophisticated reasoning will be useless. We have some evidence [10] that object detectors can work well, if they are trained on many examples of whole objects and tested on instances from the same categories. But can these methods learn to recognize parts or broad categories in a way that generalizes across categories?

Our dataset allows us to answer this question. Using the code from Felzenszwalb et al. [13] and our training set, we train detectors for parts (e.g., "leg" or "wheel"), superordinate categories (e.g., "four-legged animal" or "four-wheeled vehicle"), and basic-level categories (e.g., "dog" or "car"). These detectors model objects as mixtures of deformable "part" models. These parts are latent and without intermediate semantics. They are modeled by histograms of gradients (HOG) and allowed limited movement, providing robustness to small deformations. We use the default settings, modeling objects as a mixture of two components, each with a root and five latent parts (see [13] for further details). We find that both the mixture model and the latent "parts" improve recognition performance, even when detecting simple semantic parts such as legs or wheels. Detection is performed by sliding window, followed by non-maximum suppression with 0.5 overlap threshold for categories and 0.25 for parts. The detector SVM outputs are calibrated using Platt's probabilistic outputs algorithm [29] (fitting a sigmoid) on the training set.

In Figure 3, we show the test accuracy of the trained detectors for both familiar and unfamiliar objects. For instance, the "four-legged animal" detector needs to generalize from the familiar objects – camels, dogs, elk, lizards, and elephants – to the unfamiliar objects, such as cows, cats, and alligators. The superordinate categories tend to achieve about 60% of the recall at the same false positive rate, while part detectors have greater variation in performance. Some parts are relatively easily detected and generalized. These parts include leg, wing, head, eye, and ear for animals, and wheel, license plate, and side window for vehicles. Some parts (not shown), such as rear-view mirror were too small or too infrequent to learn well. Overall, these detection results show a surprising degree of generalization across categories. This gives us hope for more integrated object models.

2.3. Shared Correlations and Spatial Relations

Part locations, categories, and other attributes of objects are strongly correlated. Further, many of these correlations are shared across categories. While dogs and cows look quite different, they have roughly the same configuration of parts, poses, and so on. Some of these attributes can be inherited from basic categories, but we also want to localize and describe objects from unfamiliar categories. For localization, we use part and category detectors based on spatial models shared across categories. For description, we propose a simple graphical model that loosely encodes attribute correlations. The model is a form of topic model with "roots" that serve as soft clusters. These clusters summarize the visual evidence in a way that allows us to infer other attributes. For example, one cluster could correspond to four-legged animals lying down and facing left, while another might correspond to flying birds.

3. Finding and Describing Objects

Given an input image, we want to find all of the objects within known domains (here, animals and vehicles) and infer their attributes. The first step is to apply our trained detectors for parts and categories. We then obtain object candidates by accumulating votes from confident detectors. The accumulation of voting confidence provides an initial score. Finally, we perform inference over our graphical model to infer likelihood of the object attributes.

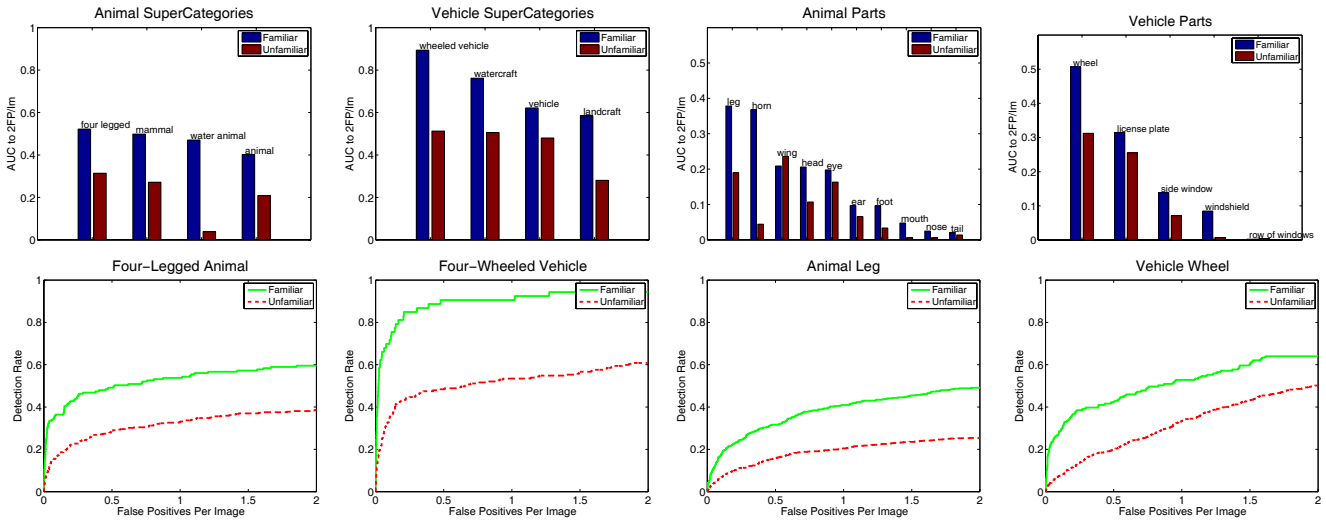
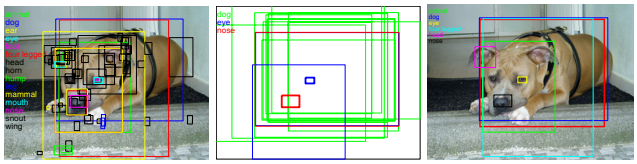


Figure 3. Current object detectors can learn parts and superordinate categories that generalize across basic-level categories. On the **top** row, we show the area under the ROC (AUC) for all detectors that are required to generalize to unfamiliar objects, with the **bottom** showing full curves for some examples. The categories of familiar objects are seen during training, while unfamiliar objects are not. AUC is computed on the curve truncated at 2FP (see Section 4.1), so that chance performance is approximately 0 and perfect is 1.



Detections Voted Boxes Candidate

Figure 4. Illustration of voting method. Confident detections vote for object position and size. Left: high confidence detections (box and name colors correspond). Center: sample of votes from three detections (thick lines are detected box, thin lines are voted boxes). Right: object candidate in red and detections that cast votes for it.

3.1. Finding Objects by Voting

We train one localizer for animals and one for vehicles that predicts the object bounding box based on the positions and confidences of category and part detections. Our voting method (illustrated in Figure 4) is strongly related to existing works that vote based on learned codewords [19, 21], distinctive keypoints [8, 38], or human parts [5]. Of these, our method is most similar to Bourdev and Malik [5], who select distinctive parts that correspond to a particular position in a supervised body pose. Our method differs in that the parts used for voting are semantic, fully supervised, and, more importantly, shared across categories.

In training, we find all correct detections above a given confidence threshold (0.01 for the calibrated detectors in our experiments). Then, we compute and store the offset in scale and position (relative to scale) for each ground truth *object* bounding box. For instance, both a detected "head" and a detected "dog" will vote for the bounding box of the entire animal. This allows us to vote from both parts and whole-object detectors. Denoting the detection box by center $\{x_d, y_d\}$ and scale $\{s_{xd}, s_{yd}\}$ and the ground truth object box $\{x_o, y_o, s_{xo}, s_{yo}\}$, the offset is $\{\frac{x_o - x_d}{s_{xd}}, \frac{y_o - y_d}{s_{yd}}, \frac{s_{xo}}{s_{xd}}, \frac{s_{yo}}{s_{yd}}\}$. During prediction, each offset gets an equal vote with the sum equal to the detection confidence; the voted box is determined by the offset and the detection bounding box. Some detectors may have hundreds of correct detections, many with nearly identical offsets. To improve efficiency, we merge nearly overlapping offsets (intersection over union threshold of 0.85) as a pre-process, accumulating votes appropriately.

During testing, we threshold detections by confidence (again at 0.01) and cast weighted votes for each offset. These need to be combined into final votes for objects. The typical procedure is accumulation through Hough voting [19, 21] or mode-finding using mean shift [38]. We found these methods difficult to work with, due to speed and the need to set various parameters. Instead, we use a simple two-step clustering procedure. The first step is to perform non-maximum suppression of voted boxes. The most confident vote is set as a cluster center. Remaining boxes in decreasing order of confidence are assigned to the existing center with highest overlap or made into a center if the maximum overlap is less than threshold (0.5). The second step is a form of k-means, iterating between computing the weighted mean of the boxes within a cluster and reassigning each box to the nearest center (using overlap, not Euclidean distance). Because these centers may drift towards each other, we repeat these two steps several times until the number of centers is left unchanged. The initial score for a candidate is given by the sum of confidences of voted boxes with at least 50% overlap.

The entire voting process takes about fifteen minutes to find all animals or vehicles in the set of 1400 test images, and it achieves high recall with few object candidates per image. With roughly 10-20 candidates per image, the system achieves 85% recall for familiar objects, and roughly 70% recall for unfamiliar objects. The parts improve recall, es-

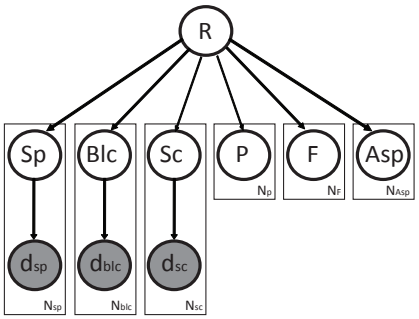


Figure 5. Graphical model representation of the root model. "R" is the root node through which the attribute nodes "Sp", "Blc", "Sc", "P", "F", and "Asp" communicate. Shaded nodes are observed detector responses for spatial parts, basic level categories and superordinate categories.

pecially for unfamiliar animals: without them, recall drops by about 15%.

Though they improve recall, the part detections add little weight to the voting score because they are not independently confident. To make better use of them, we rescore the detections by training logistic regression on the voting score and the localized part and category detections that are described next.

3.2. Describing Localized Objects

The description task aims to predict binary attributes of a given localized object. We do so by performing inference on the graphical model presented in Figure 5. In our model, the "root" node generates each of the attributes, some of which generate detector observations. The *spatial part* (Sp) nodes encode the visibility of the parts in one of the six spatial bins (whole, top, bottom, left, center, right). The d_{sp} encodes the strongest detector response in each of the spatial bins. BLC stands for the *basic level categories*. The d_{blc} is the maximum detector response with sufficient overlap with the region of interest. *Superordinate categories* are handled by the Sc node. Similar to the d_{blc}, the d_{sc} is the maximum detector response for superordinate categories. The remaining nodes encode attributes which do not directly rely on any detector. These attributes may not be visually obvious, such as functional attributes (F), hard to predict directly, such as aspect (Asp), or not have enough training examples to train appearance models. The node "P" indicates if an object has an attribute or not. This is different from the visibility of an attribute. For instance, "dog" has "leg" regardless of the "leg" being visible or not. For this purpose, we consider including a set of nodes "Sp" for spatial visible parts and another set of nodes "P" to consider the potentials of having a part. We also have nodes for the functional attributes of objects such as "Can this object bite?"

The model is learned using Expectation Maximization (EM). As the nodes are multinomial, the derivation is straightforward. We show that this model improves the attribute prediction for familiar objects and that of unfamiliar objects by considerable margins (Table 2).

Inference can be done in closed form shown in Equations 1 and 2 by marginalizing over attributes with and without learned detectors. Equation 1 computes the marginals given the observations for attributes A_i for which we have learned detectors ($A_i \in \{Sp, Blc, Sc\}$). $A_j \neq A_i$ corresponds to all other nodes for which we have detectors.

$$P(A_i = a_i|\bar{d}) \propto \sum_R P(R)P(a_i|R)\frac{P(a_i|d_i)}{P(a_i)} * \prod_{A_{j\neq i}} \sum_{A_{j\neq i}} P(A_j|R)\frac{P(A_j|d_j)}{P(A_j)} \quad (1)$$

The inference on attributes $B_i \in \{P, F, Asp\}$ without any learned detector is obtained by Equation 2:

$$P(B_i = b_i|\bar{d}) \propto \sum_R P(R)P(b_i|R)\prod_{A_j}\sum_{A_j} P(A_j|R)\frac{P(A_j|d_j)}{P(A_j)} \quad (2)$$

where A_j corresponds to all the nodes with detectors.
This framework allows us to learn separate root models for each domain and perform inference over them jointly. To do so, we can simply change the priors for each root to sum to one over all domains.

4. Experiments

We perform experiments on two tasks: (1) find *all* animals or vehicles; and (2) describe localized objects by their attributes. In each case, we measure how well we perform for familiar objects and for the unfamiliar objects. In all, our experiments show that part and superordinate detectors can generalize across basic categories (Figure 3) and that modeling objects in terms of shared properties allows much better localization (Figures 6, 7) and description (Table 2, Figure 8) for both familiar and unfamiliar objects.

4.1. Experimental Setup

Baseline. Our baseline uses top-notch detectors [13] to learn basic-level categories. For localization, we calibrate the detectors and perform non-maximum suppression. For description, we model the attributes as probabilistically inherited from the categories, and infer them by marginalization. Essentially, the baseline makes the basic categories the roots of our model and does not use additional detectors. This is similar to the experiments in [18]. Our method substantially outperforms the baseline in both tasks, especially localization and prediction of pose.

Evaluation. To evaluate localization, we use area under the ROC curve, truncated at 2FP per image to emphasize the high precision portion. In contrast with average precision, a recently popular performance measure [10], our measure

Localization AUC	Animal			Vehicle		
	F	U	C	F	U	C
BLC Baseline	.364	.126	.203	.644	.313	.425
Voting	.456	.230	.303	.679	.441	.521
Full Model	**.471**	**.247**	**.320**	**.678**	**.468**	**.539**

Table 1. We compare AUC for localizing familiar and unfamiliar objects (F=familiar, U=unfamiliar, C=combined) to a baseline that uses a detector trained only on basic categories.

does not depend strongly on the density of positive examples. This is important because it allows us to meaningfully compare curves computed for different populations of objects.

To evaluate the description task, we compute an ROC curve and its area (AUC) for each attribute. We then average the AUCs within each of these attribute types: basic-level category, superordinate category, existence of parts, pose, and function (section 3.2).

4.2. Results for Finding Objects

In Table 1 and Figure 6, we compare our ability to find animals or vehicles to a baseline. We can draw two conclusions. First, we are better able to recover familiar objects than unfamiliar objects, as expected, we also do well on unfamiliar objects (Figure 7). Second, our method outperforms the baseline by a large margin, especially for unfamiliar objects. The improvement is amazing considering that we use the same detection method as the baseline and see the same training examples. The difference is due to our appropriate use of our shared part and superordinate detectors.

The baseline is computed by performing non-maximum suppression over the calibrated basic-level category (BLC) detectors. Our method votes for object candidates using part and category detection, weighted by confidence, and sums the votes as a score (Section 3.1). We also tried our voting method using only BLC detectors and achieved similar results to the baseline, ensuring that the improvement is due to our shared parts and superordinate categories. Although part detectors improve our recall, they do not have large impact on the voting score because they are rarely confident. To make better use of them, we re-rank top candidates using logistic regression on the voting score and all detector responses in our root model (d_{sp}, d_{blc}, d_{sc} in Figure 5) . This provides a small but significant improvement.

4.3. Results for Describing Objects

Table 2 compares the ability of the root model for predicting attributes of objects with that of using the baseline. Table 2 shows the area under ROC curve for familiar and unfamiliar objects. Our method improves attribute predictions for familiar and unfamiliar objects by large margins. These results show the ability of this model to generalize across categories. Figure 8 depicts qualitative object description results.

5. Discussion

We have shown how to learn more accurate and detailed object models through shared representations. Our models are simple, and we rely on no hand-tuned parameters. Carefully designed representations and a small amount of additional annotation is sufficient to achieve quantitatively and qualitatively better results than existing detectors that are given only the names of objects. Our study is only recently made possible by advances in detection, prolific data, and large-scale annotation services. We believe that our dataset will open new avenues for familiar problems. For example, active learning methods can use intermediate semantics to form more detailed queries. Contextual methods may prove especially helpful for superordinate categories. There is also much room to better model the correlations of attributes, encoding prior knowledge as soft constraints. Computer vision has long progressed by subdividing problems; now we can further progress by rebuilding.

6. Acknowledgments

This work was supported in part by a Google Research Award and the National Science Foundation under IIS-0904209, 0534837, and 0803603, and in part by the Office of Naval Research under N00014-01-1-0890 as part of the MURI program. Ali Farhadi was supported by Google Ph.D fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of NSF, Google, or ONR. The authors would like to thank Alexander Sorokin for insightful discussions on Mechanical Turk.

References

[1] N. Ahuja and S. Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *ICCV*, 2007.

[2] E. Bart and S. Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, 2005.

[3] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Mach.* MIT Press, 2007.

[4] I. Bierderman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

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

[6] R. Brooks, R. Greiner, and T. Binford. Model-based three-dimensional interpretation of two-dimensional images. In *IJCAI*, 1979.

[7] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[8] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

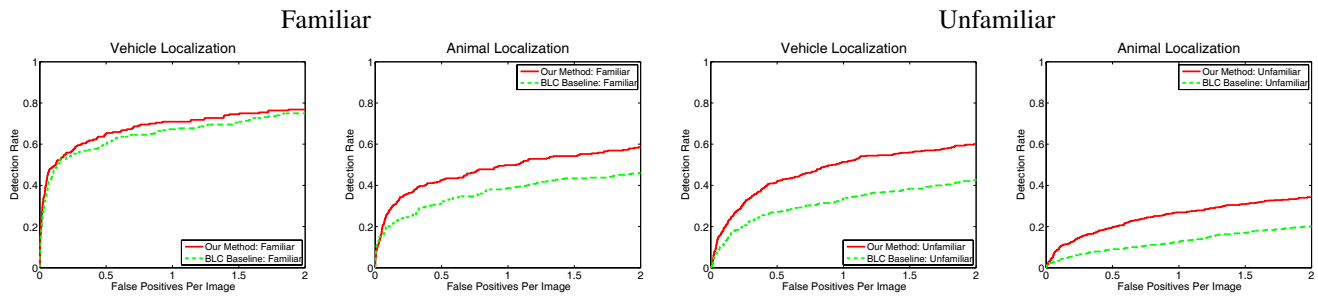


Figure 6. We compare our ability to detect familiar and unfamiliar animals and vehicles. Our model integrates shared parts and superordinate detectors. The baseline uses only the standard basic-level detectors.

Figure 7. Our system can find animals and vehicles and localize their parts, even if it has never seen them before. The first two rows show examples of detections for animals and vehicles like cows, cats, bicycles, buses, carriages, or horses, that our system never observed. Often times, detected parts help the final detection of unfamiliar objects. For example the the legs of the cows in the first row, or the wheels of the carriage in the second row. Each solid-line bounding box is an object detection above a given threshold (**red=vehicle, green=animal**), and dashed boxes show part detections that helped to find the object (first three letters of part name shown). Black boxes indicate detected categories. The third row depicts examples of detecting familiar objects. The bottom row shows examples of mistakes our system makes. For example predicting vehicle wings for the motorcycle. This is mainly because of the fact that we do not enforce any consistency during localization.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[12] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *CVPR*, 2007.

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[14] A. Guzman. Computer recognition of three-dimensional objects in a visual scene. Technical Report MAC-TR-59, MIT, 1968.

Domain	Method	Average		Has Part		Basic-Cat		Super-Cat		Function		Pose	
		F	UnF	F	UnF	F	UnF	F	UnF	F	UnF	F	UnF
Animal	Root Model	0.757	0.646	0.798	0.747	0.755	NA	0.761	0.591	0.807	0.602	0.665	0.649
	Baseline	0.701	0.591	0.770	0.648	0.721	NA	0.710	0.618	0.732	0.567	0.571	0.532
Vehicle	Root Model	0.854	0.700	0.929	0.752	0.885	NA	0.891	0.778	0.922	0.691	0.643	0.578
	Baseline	0.781	0.652	0.870	0.723	0.841	NA	0.849	0.717	0.801	0.637	0.544	0.533

Table 2. We compare our ability to infer attributes to a baseline that infers them directly from predicted basic-level categories (measured by average AUC). From part and category detectors, our method infers what parts an object has (regardless of its visibility), the basic-level category (e.g., "dog"), the superordinate-category ("four-legged animal"), function ("can jump"), and pose or viewpoint ("lying down", "front side"). Our model results in higher accuracy for each type of attribute, despite that some are directly predictable from basic categories. **F**amiliar objects are seen during training. **Unf**amiliar objects are not.

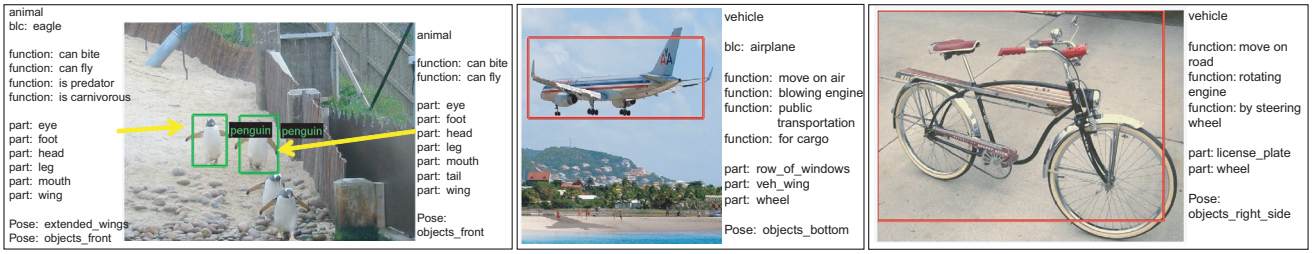


Figure 8. Our system can describe localized objects in terms of their attributes which are inferred from localized parts and categories. The list of the attributes here are all of the attributes above some thresholds. Our system correctly predicts the extended wing pose of one penguin, the bottom view of the airplane, and the right side view of the bicycle. Penguin and airplane are familiar and bicycle is an unfamiliar object for our system.

[15] A. Hanson and E. Riseman. VISIONS: A computer system for interpreting scenes. In *Computer Vision Systems*, 1978.

[16] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006.

[17] N. Kumar, A. Berg, P. Belhumeur, and S. K. Nayar. Attribute and simile classifiesrs for face verification. In *ICCV*, 2009.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1–3):259–289.

[20] F. Li, R. Fergus, , and P. Perona. One-shot learning of object categories. *PAMI*, 28(4):594–611.

[21] S. Maji and J. Malik. Object detection using a max-margin hough tranform. In *CVPR*, 2009.

[22] M. Marszałek and C. Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, 2008.

[23] K. McRae, G. Cree, M. Seidenberg, and C. McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559, 2005.

[24] G. L. Murphy. *The Big Book of Concepts*. The MIT Press, March 2002.

[25] Y. Ohta. *Knowledge-Based Interpretation Of Outdoor Natural Color Scenes*. Pitman, 1985.

[26] Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *IJCPR*, pages 752–754, 1978.

[27] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *CVPR*, 2006.

[28] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes.

[29] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[30] D. Rakison, , and L. M. Oaks. *Early Category and Concept Development*. Oxford University Press, Oxford Oxfordshire, 2003.

[31] E. Rivlin, S. J. Dickinson, and A. Rosenfeld. Object recognition by functional parts. In *Image Understanding Workshop*, 1994.

[32] L. Roberts. Machine perception of 3-D solids. In *OEOIP*, pages 159–197, 1965.

[33] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. B. Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, July 1976.

[34] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. 2008.

[35] M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In *ICCV*, 2009.

[36] S. Thrun. *Explanation-Based Neural Network Learning: A Lifelong Learning Approach*. Kluwer Academic Publishers, Boston, MA, 1996.

[37] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*. 2004.

[38] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[39] T. F. Wu, G. S. Xia, and S. C. Zhu. Compositional boosting for computing hierarchical image structures. In *CVPR*, 2007.

[40] S.-C. Zhu and D. Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.