# Birdlets: Subordinate Categorization Using Volumetric Primitives and Pose-Normalized Appearance

Ryan Farrell[1,2], Om Oza[1], Ning Zhang[2], Vlad I. Morariu[1], Trevor Darrell[2], Larry S. Davis[1]
[1]University of Maryland, College Park [2]University of California, Berkeley
{farrell,nzhang,trevor}@eecs.berkeley.edu,{omoza,morariu,lsd}@umiacs.umd.edu

## Abstract

*Subordinate-level categorization typically rests on establishing salient distinctions between part-level characteristics of objects, in contrast to basic-level categorization, where the presence or absence of parts is determinative. We develop an approach for subordinate categorization in vision, focusing on an avian domain due to the fine-grained structure of the category taxonomy for this domain. We explore a pose-normalized appearance model based on a volumetric poselet scheme. The variation in shape and appearance properties of these parts across a taxonomy provides the cues needed for subordinate categorization. Training pose detectors requires a relatively large amount of training data per category when done from scratch; using a subordinate-level approach, we exploit a pose classifier trained at the basic-level, and extract part appearance and shape information to build subordinate-level models. Our model associates the underlying image pattern parameters used for detection with corresponding volumetric part location, scale and orientation parameters. These parameters implicitly define a mapping from the image pixels into a pose-normalized appearance space, removing view and pose dependencies, facilitating fine-grained categorization from relatively few training examples.*

## 1. Introduction

In recent years, the computer vision community has devoted extensive efforts toward the development of computational techniques for object recognition. These efforts, however, have focused almost exclusively on the recognition of basic-level categories; relatively few have addressed the broad continuum of fine-grained or subordinate categories which lies between the two extremes of individuals (e.g. face recognition, biometrics) and basic-level categories (e.g. Caltech-256 *etc.*).

In cognitive psychology, Rosch *et al*. [43] proposed that, whereas basic-level categories are principally defined by their parts, subordinate level categories are distinguished by the differing properties of these parts. This theory suggests that the capacity to differentiate subordinate categories hinges not only on the successful recognition of individual parts but, perhaps more particularly upon understanding how these part "properties" vary across subordinate categories. While recent advances on part-based and attribute-based recognition are promising, general and view-independent identification of part-specific attributes in novel images remains somewhat elusive.

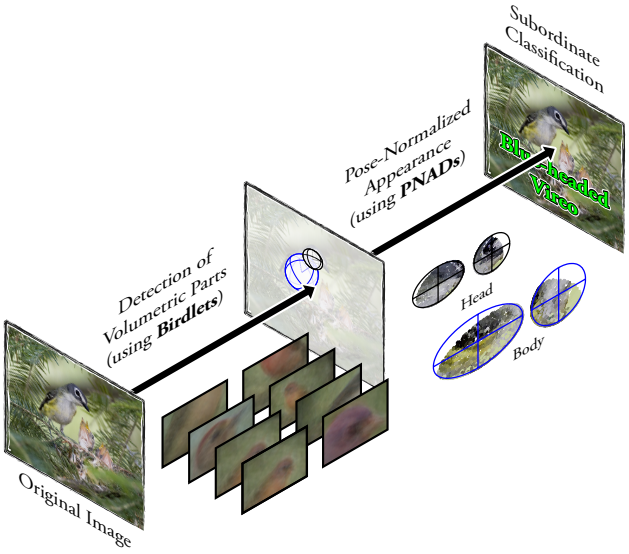



Figure 1. **Overview of the Proposed Approach.** Basic-level categories are modeled by a configuration of volumetric primitives or parts. Detection recovers these parts and enables application of a pose-normalized appearance model for classification within a taxonomy of subordinate categories.

We tackle the problem of subordinate categorization, proposing a solution that simultaneously addresses the challenges of localizing and describing the class-defining parts. Our approach (see Figure 1) builds upon the Poselet detection framework recently proposed by Bourdev *et al*. [7, 8]. The strength that we see in this framework is that, in theory,

the model allows for specific types of training annotations to be recovered from detections in test images. Our approach is also motivated by Biederman's theory of non-accidental arrangements of geometric primitives [5, 6]. We use a simple configuration of volumetric primitives to represent the basic-level class. Then, following Rosch *et al*., variations in the shape, configuration and appearance of these volumetric parts provide the basis for subordinate discrimination.

Our proposed approach contributes three main innovations:

*(i)* a framework, based on Poselets, for detecting volumetric part models, used both to find the basic-level object and to convey information about part shape and configuration;

*(ii)* a pose-normalized appearance model (similar to representations such as Active Appearance Models [13] and Morphable Models [33] used in the domain of faces) which is used to effectively compare part appearances in a test image to those of subordinate category training examples; and

*(iii)* a classification model, based on Stacked Evidence Trees [39], which aggregates information about part properties (shape, configuration and appearance) and leverages the underlying taxonomy.

We demonstrate experimentally that the proposed approach enhances the performance for view-independent recognition of subordinate categories.

## 2. Related Work

The problem of subordinate categorization has been previously examined. Hillel *et al*. [2] performed experiments on two subclasses for each of six basic categories (e.g. Grand vs. Upright Pianos). Nilsback and Zisserman [41, 42] considered subordinate categories of flowers (introducing the 17- and 102-category Oxford Flowers datasets), whereas Martínez-Muñoz *et al*. [39] considered subordinate categorization of stonefly larvae, a domain which exhibits tremendous visual similarity. These approaches focused primarily on discriminative learning of image features, an approach that does not generalize for view-independent categorization of part-based objects that exhibit significant pose variation.

There are various methods that have been proposed for learning part-based object representations. Constellation models [11, 48] and their computationally attractive variants [14, 27] are composed of a set of local part detectors together with one or more probability density functions describing the parts' relative locations. Felzenszwalb and Huttenlocher [24] proposed an efficient framework implementing Fischler and Elschlager's Pictorial Structure model [29], which represents an object by a collection of parts, interconnected as if by elastic springs. This Deformable Part Model has culminated in Felzenszwalb *et al*.'s recent work using Latent SVMs [23] to discriminatively train class-specific object detectors. Ferrari among others have explored the use of contours in object representation [28]. While these models perform well for objects that exhibit minimal articulation or pose variation, they are unsatisfactory for objects with high intra-class variability or significant articulation.

There is also a growing body of work that seeks to leverage similarities between categories to improve recognition performance. We consider two principal areas of interest: first, class taxonomies or hierarchies and, second, attribute-based models. Unsupervised hierarchical approaches range from constructing latent topic hierarchies [3] to sharing classifiers [1] or visual parts [45] to constructing efficient classification trees [31, 38]. Each such approach provides insights or advances toward efficiently solving basic-level classification. These unsupervised approaches, however, cannot be readily applied to the problem of distinguishing closely-related subordinate categories which, by definition, share a common set of parts and yet can have both subtle and drastic appearance variation.

Techniques that leverage the semantic class hierarchy should possess an inherent advantage over those that do not. Supervised methods that utilize such information (as contained in WordNet for example) include the sharing of training examples across semantically similar categories [26] and combining information from different levels of the semantic hierarchy [50]. Deng *et al*. [16] consider exploiting the semantic hierarchy in the context of more than 10,000 categories (using the ImageNet [17] dataset).

A growing interesting in attribute-based recognition has produced some notable advances. Representative work in this area includes Farhadi *et al*. [19, 20], Kumar *et al*. [34] Lampert *et al*. [35] and Wang and Forsyth [47]. These techniques often learn discriminative models from attribute-labeled training data and subsequently apply the learnt models to estimate the appropriate visual attributes present in a test image. Attribute-based models are particularly well-suited for addressing the one-shot learning problem (previously considered in [21, 22, 25, 40] among others). Note that while these approaches are effective for the recovery of object level attributes such as brown, furry, spotted and even four-legged, they are generally insufficient to model subtle differences between parts necessary for subordinate categorization.

An interesting exception is the innovative work of Branson *et al*. [9] which proposes improving recognition accuracy by interleaving computation with attribute queries made to a human subject. This method performs effective, though not automatic, recognition in a large, 200-category bird dataset [49][1]. Additionally, in the context of subordi-

[1] Additional details on the CUB-200 dataset can be found in Section 6 which describes our experiments.

nate categorization, the attribute-based work of Berg *et al*. [4] is also of interest as it attempts to discover (and localize) visual attributes which can be used to differentiate classes within a basic-level category (e.g. stiletto, running shoe, sandal, *etc*.). This approach is somewhat limited, however, in that its training data is segmented from any background and also must be in a similar pose/orientation.

Before proceeding to describe our approach, we first visit the theory initially put forth by Marr and Nishihara [37] and later extended with Biederman's geons [5] which suggests that object perception is largely governed by recognition of three-dimensional parts in particular configurations. While subsequent research has questioned certain aspects such as view invariance [44], this theory of perception as the search for arrangements of non-accidental structures has survived. Biederman *et al*. revisited it in the specific context of subordinate-level classification [6]. This theory provides support for the proposed approach which models a basic-level category with geometric primitives, and then couples the statistical variation of the parts' shape and arrangement with their appearance to represent subordinate classes.

## 3. Subordinate Categorization in an Avian Domain

We begin by considering more closely the problem of subordinate categorization, highlighting some of the ways it differs from basic-level categorization. The seminal work of Rosch *et al*. [43] provided experimental evidence in support of a distinction between levels of abstraction within a taxonomy: superordinate, basic, and subordinate (in decreasing order of inclusivity). Rosch *et al*. contend that basic-level categories generally possess the highest cue validity $P(category|cue)$, as superordinate-level categories, being more inclusive, have fewer attributes in common and subordinate-level categories share most of their attributes with contrasting subordinate categories.

### 3.1. Basic- and Subordinate-Level Categorization

Objects within a superordinate category tend to share common material and/or functional properties (sensory-motor "affordances" to use Gibson's terminology [30]). In contrast, a (and perhaps the) key characteristic of categories at the basic-level is shape. Rosch *et al*. include in their definition of shape "the structural relationship of the parts of an object to each other - for example, the visual representation of the legs, seat, and back of a chair and of the way in which those parts of the chair are placed in relation to one another."

This notion of basic-level shape as a fixed set of parts in an expected arrangement agrees strongly with Biederman's theory of Recognition-by-Components [5] which suggests that a category may be represented by volumetric components or primitives called "geons" (blocks, cylinders, cones, *etc*.) in a particular configuration. While Biederman's theory presents a broad perspective on the human recognition process (edge extraction and parsing, identification of components, matching to known configurations, object identification), we focus on this underlying representation of basic-level categories: a configuration of volumetric parts.

This basic-level representation is intuitive for many natural categories. Objects within a category (dogs or trees, for example) share a common set of parts in a more-or-less prototypical configuration ($\{$head, body, legs, tail$\}$ and $\{$trunk, branches, leaf canopy$\}$ respectively). Within each such category, the configuration and "connectivity" of these parts is generally highly-constrained.

Differentiation amongst subordinate categories (*e.g*. between sports cars and sedans or even between different brands/models), however, must rely on more than simply the presence and/or configuration of these parts. We thus consider *properties* of these parts, both quantitative properties such as shape variation (aspect, relative size) or structural relationships (relative position/angle) and qualitative appearance properties such as color, material and texture.

We have selected birds as the domain for our experimental evaluation for a variety of reasons. There are several basic-level categories for which vision datasets include many subordinate classes. To our knowledge, none is larger than the recently introduced Caltech/UCSD Birds dataset (CUB-200) [49] which includes 200 distinct avian species. While some categories are readily identified by their unique shape, pose, or appearance, the distinctions between other categories are very subtle. Due to highly variable appearances and articulation, birds are also extremely challenging to even detect, consistently the most difficult across the 20 categories on the Pascal VOC challenge [18]. Ultimately, however, our decision to use birds as a domain in which to explore subordinate visual categorization is principally motivated by their suitability for our pose-normalized representation.

### 3.2. Pose-Normalized Appearance Representation

Following Rosch's prototype theory, we distinguish subordinate categories based on the geometric shape and photometric appearance properties of their basic-level parts. In describing our appearance representation, we begin with a basic-level object, represented as a constellation of volumetric parts. The detection process provides estimates for each part's respective parameters: location, scale and orientation. The geometric shape and arrangement properties can be used to influence categorization. Within the domain of birds, taxonomic guidance by shape is intuitive; individuals with minimal expertise in recognizing birds can correctly assign a silhouette to its respective family (*e.g*. duck, heron, hawk, owl, songbird, *etc*.).

As far as the volumetric part appearance properties, the

primary difficulty is pose variation relative to the camera, an issue that complicates the comparison of part appearances observed from different angles. To overcome this challenge, we propose a pose-normalization approach leveraging the detected volumetric parts. Fundamental to our approach, this technique imposes a surface parameterization on the volumetric part, the parameterization serving as a basis for a non-parametric appearance representation. Comparisons between images are made not in image space, but on a distribution of patch descriptors in the parameterized space of estimated surface normals.

In our part model, we have two ellipsoids, one for the head and one for the body. For a given ellipsoid, we use the pose parameters: ellipsoid center (x,y), scale (cross-section and axial aspect ratio) and orientation (represented as a quaternion), to determine the transformation that maps points on a unit sphere onto the ellipsoid's surface. The inverse of this transformation allows us to map image points (those within the ellipsoid's silhouette) back onto the unit sphere. Instead of parameterizing in the sphere's space, we randomly sample points on the sphere, transform them onto the ellipse's surface and compute their normals (using the inverse transform), ensuring that they are visible (facing the camera). As depicted in Figure 2, for each such point on the ellipsoid's surface, we find the *tangent patch*, a small square patch on the tangent plane centered at the point. The corners of the tangent patch are projected back into the image forming a parallelogram. The parallelogram's pixel contents are warped onto the square fronto-parallel tangent patch (purple in Figure 2) from which local appearance features are derived (we use a color-SIFT descriptor aligned with the dominant gradient orientation). We couple each patch's location and appearance by concatenating the normal vector (blue in Figure 2) onto the extracted appearance descriptor (red in Figure 2), yielding a pose-normalized appearance descriptor, or PNAD. After sampling several such points/patches, we accumulate a non-parametric representation for the visible portion of the ellipsoidal part.

## 4. Volumetric Object Localization

As suggested in the introduction, the primary requirement for successful differentiation of subordinate categories is an ability to find parts and understand how these parts vary (or alternatively, how the "properties" of these parts vary) across different subordinate categories. To address the problems of localizing and describing the class-defining parts simultaneously, we adopt the Poselet framework recently proposed by Bourdev *et al*. [7, 8], using an object model comprised of volumetric primitives instead of 2D or 3D keypoints. We provide a brief description of the approach while highlighting changes needed for our volumetric implementation.

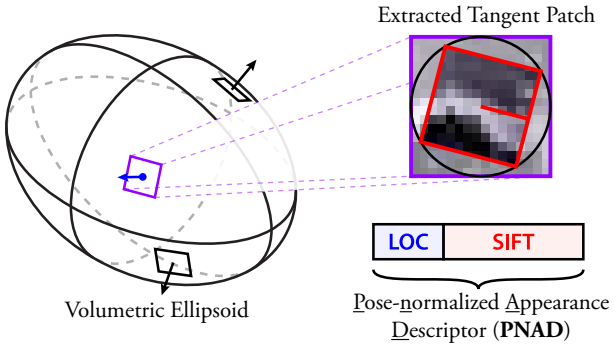



Figure 2. **Pose-normalized Appearance Descriptor (PNAD).** For each ellipsoidal part, tangent patches (purple) with corresponding appearance descriptors (red) are extracted at sampled points (blue normal vector) and a Pose-normalized Appearance Descriptor, or PNAD, is formed by concatenating the location and appearance information.

### 4.1. Birdlets: Volumetric Primitive Templates

While the Poselet framework represents a basic-level category as a constellation of 2D keypoints, our approach creates "Birdlets", templates based instead on solid volumetric primitives, consistent with Biederman's notion of basic-level categories as arrangements of 3D geometric primitives. Where the former technique estimates the image location of each keypoint, the utility of using volumetric parts lies in its potential to estimate various geometric quantities for each of the volumetric elements that collectively comprise the basic-level category model. Examples of such geometric attributes (or "properties") include part location, size/aspect, and orientation, and can encode intrinsic category characteristics such as the cross-section or aspect of a bird's body relative to the size of its head.

This volumetric model is particularly well suited for birds, as the avian counterparts for interior mammalian joints (*e.g*. shoulders, elbows, hips, knees) are often obscured by a bird's plumage and are thus very difficult to specify in a typical image. Moreover, the surface or skeletal keypoints used in the original Poselet models capture part proportions (e.g. cross-section, aspect) poorly. The proposed model, therefore, includes visible point features (beaktip, eyes, wingtips, feet, and tail) only to assist in configuration alignment; the model remains focused, however, on its two volumetric components. The bird's head and body are each represented by prolate ellipsoids (a sphere stretched along one axis)with 7 parameters: image location (x, y), 3D-orientation (a 3-DOF quaternion), and scale (circular cross-section and axial length). Where one could try to model a bird with additional primitives, this simplified version (or "partial version" as Biederman calls it [5, p. 131]) captures the essence of shape and enables the pose-normalized appearance representation.

### 4.2. Training and Detection

The Poselet framework requires images annotated with configuration landmarks (the 2D or 3D keypoint locations in Bourdev *et al*.; the location, orientation and scale of volumetric primitives in our case). These annotations serve to help find training examples that share similar local pose or configuration (the entire pose need not be similar, just the part(s) or keypoints in question). In this manner, images depicting similar poses relative to the camera are grouped together.

Birdlet training takes a certain *base* training image and determines a *selection window* overlapping some subset of the volumetric parts (in our case, this could be the head, the body or both). Next, the *pose distance* to each of the other training images is computed, based on the similarity in parameters for this subset of parts (*i.e*. can the two images be registered to one another such that the parts align well). Specifically, this distance is computed using terms for rotation (geodesic distance on 4D surface of quaternion rotations), scale (computed on cross-section and aspect after scaling to equal volume) and translation (generally ignored as single ellipsoids can be brought into precise alignment as can the dipoles formed by ellipsoid centers)

The $n-1$ closest training images are selected (we nominally use $n = 50$) and the similarity transform to align each to the base image is determined. With this transform, the parts now line up (as best as can be done with the 2D similarity transformation) and the corresponding image features should now be well aligned also. Now, for each of $n$ training images (the base and the $n-1$ closest in terms of pose distance) which have been brought into alignment, the pixels in the selected window are mapped into a canonical rectangular *patch* ($96 \times 64$ in our case) and a HOG vector [15] is extracted (the concatenation of HOG features across $8 \times 8$ blocks). These $n$ HOG vectors are used as positive examples, together with a much larger set of negative HOG vectors (extracted from other random windows in the training data), are used to train an SVM classifier to discriminate this birdlet from background patterns. Like Bourdev *et al*., we use a retraining stage, collecting false positives predicted by the initial classifier and feeding these as additional negative examples in order to train the final classifier for this birdlet.

For detection, our birdlet classifier will evaluate patches in a test image using a sliding window (scanning over locations and scales), responding with a probability of how similar each scanned patch appears to the positive examples that the classifier was trained with. Windows with high response probabilities are labeled as *activations* for the given birdlet.

The great benefit that we saw in the framework of Bourdev *et al*. is that the birdlets we train facilitate detection, but moreover provide information about the pose or part-configuration. A birdlet activation provides an estimate or vote toward the parameters of those volumetric parts that overlapped the birdlet's selection window. Hence, whereas other techniques typically learn a model on latent parts, the birdlet model maps the image patterns within the selection window to the semantically meaningful volumetric primitives, inherently providing a level of visual correspondence across instances (and views).

Many such birdlet templates are trained, binding images cues from the training set with their counterpart volumetric part annotations. The collection of birdlets is then applied to a test image producing a set of birdlet activations. Each activation has an associated probability (derived from the corresponding classifier's response) as well as the distribution on part parameters it acquired during training (this distribution is a simple tabulation on the parameters of the overlapping parts once aligned). The birdlet normalizes the distribution relative to the height of the patch, such that for a given activation window, the normalized location and relative size information can be scaled up the activation window, thus converting it to a prediction in the test image. Our implementation uses a non-parametric (kernel density estimate) density to represent each ellipsoids 7-D parameter space.

The final step is to cluster the set of activations into one or more *final detections* with the corresponding volumetric part estimates. The approach that we have taken for this clustering is to compute the pairwise consistency of activation, determined by symmetric K-L divergence between the parameter distributions of the corresponding parts shared by the activations' respective birdlets. We take the pair of activations with the highest consistency (and activation probability or response) and draw the volumetric parts' parameters from their distributions. In theory we can sample from the combined distribution, however, in practice, we found it effective to predict the parameters of each birdlet's base training image (for some birdlets, there are small clusters of examples with similar pose, and thus only a few training examples that share similar parameters).

## 5. Integrated Classification

Our approach uses an integrated classification technique based on Stacked Evidence Trees model proposed by Martinez-Muñoz *et al*. [39]. The authors describe this approach as an alternative to dictionary learning, being instead a way of "discriminatively structuring the evidence in the training set". This model relies on a Random Forest [10] constructed such that all leaf nodes of the constituent random trees are required to have a specified minimum number (*e.g*. 20) of training samples. In this manner, when a query sample is passed through a random tree and reaches a particular leaf node, the tree returns the distribution across class labels corresponding to training examples that reached that node. For a given image, features are extracted densely. As these features are dropped through the trained random

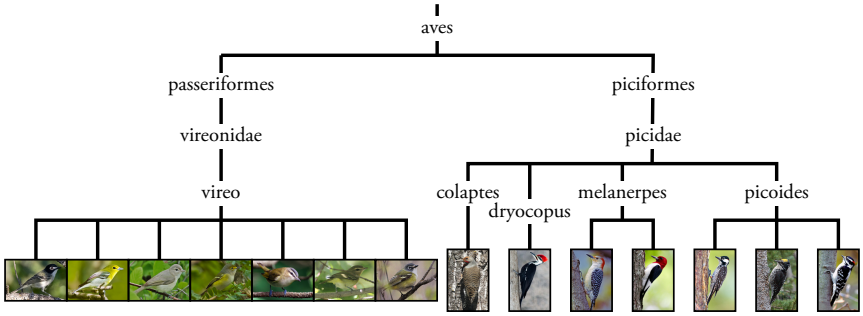



Figure 3. **Partial Taxonomy for CUB-200.** Two family subset (vireos and woodpeckers) from the CUB-200 Dataset.

forest, the class label distribution vectors are collected and aggregated into an "evidence" vector, each feature effectively voting for the category of the image. A second-stage ("stacked") multiclass adaboost classifier is then applied to the class distribution evidence vector, producing the final category prediction.

The Stacked Evidence Trees model was selected principally for the way that it complements the Pose-Normalized Appearance model, providing an attractive solution to the problem of varying surface visibility. In general, a volumetric primitive has only half of its surface facing the camera, the remaining half is not visible. As the visible/occluded portions are different for each image (*e.g.* a bird facing the camera vs. facing left vs. facing right), it is desirable not only to map the visible portions into a common (pose-normalized) space, but moreover, to effectively mask which part(s) of this common space should be used for classifying each given image.

As described earlier, the Pose-Normalized Appearance space allows us to compare corresponding parts. Specifically, a PNAD (Pose-Normalized Appearance Descriptor) feature couples local appearance information with parameterized surface location. However, due to the issue of feature visibility, one cannot simply quantize this joint appearance/surface location space and use a bag-of-words approach for classification. The Stacked Evidence Tree on the other hand becomes a highly-efficient retrieval tool, taking a test feature and finding a set of training features (namely those in the corresponding leaf nodes) that are similar both in appearance and surface location, and ultimately returning the class label distribution across this similar set.

An appealing characteristic of the Stacked classifier is the ability to combine multiple feature types by merely concatenating various evidence. In our case, we view this as the means to combine part appearance (PNADs) together with other potential sources of discriminative information. We consider combining shape and arrangement parameters (*e.g.* part cross-section/aspect, relative sizes/orientations between parts, *etc.*) as well as taxonomic training data.

One additional potential source of information which we are not currently using is the birdlet activations that contributed to the detection. When a given birdlet is trained, the other examples selected as positive patches (based on similar configuration) may collectively convey information at test time about the category of detections involving a high-probability activation of the birdlet in question.

## 6. Experimental Results

Now that we have described detection of volumetric primitives, pose-normalized appearance representation, and integrated classification, we present some experiments in support of this framework.

### 6.1. Dataset, Implementation Details, *etc*.

First utilized by Branson *et al*. [9], the Caltech-UCSD Birds 200 dataset [49] currently offers the largest number of subordinate categories for a single basic-level category. We organized the entire dataset into its proper taxonomic hierarchy (order, family, genus, species) and then selected two families to fully annotate with both 2D keypoints and 3D volumetric primitives (ellipsoids), the vireo and woodpecker families These annotations, together with near-duplicate groupings (so that near-duplicates do not straddle test-training splits), will be made publicly available to other researchers. While many annotation tasks are well-suited to crowdsourcing, we felt that proper annotation of the ellipsoids was non-trivial and accordingly have a smaller dataset than would be desirable.

As the authors of [8, 7] have only released their code for detection with a pre-trained human detection model, we had to reimplement the extensive Poselet framework in its entirety. In our birdlet implementation, we utilized LIBSVM [12] together in conjunction with Platt's algorithm [36] for converting SVM scores to probabilities. The random forest used for integrated classification was adapted from the Weka [32] machine learning package.

### 6.2. Volumetric Part Localization

Before we can consider our primary objective of subordinate categorization, we evaluate the detection of our vol-

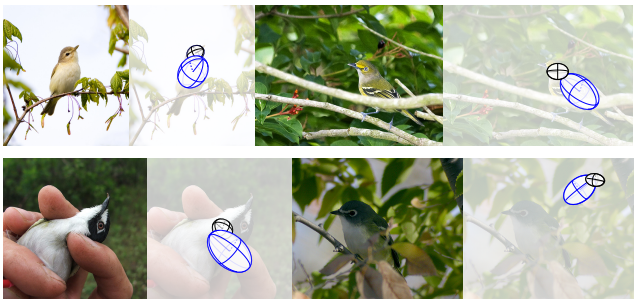


Figure 5. **Example Volumetric Primitive Detections.** Here are four representative detections. In the top two images, the bird is detected and localized with reasonable accuracy. The images in the lower row depict false positive detections, however. In the first image, a finger is incorrectly interpreted as the bird's body; the second is typical of false detections at the incorrect scale and location.

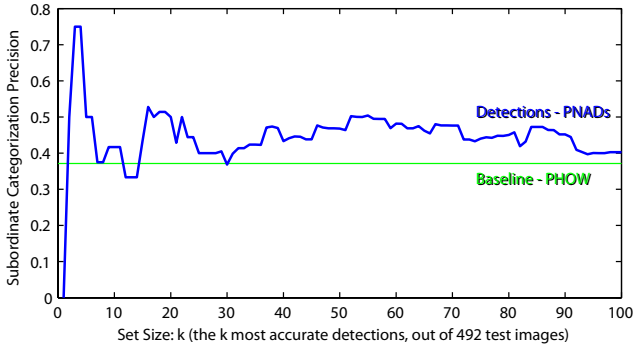



Figure 6. **Classification of Volumetric Detections.** For the $k$ top-ranked detections, this plots the corresponding PNAD-RF classification performance (using mean-average precision).

umetric part model. To train the birdlet model, we used a training split that included 15 images of each category (together with their mirrored annotations) for a total of 420 training images/annotations. The resulting birdlets (we train a set of 100 birdlets) are applied toward detection on the remaining 492 test images.

Some examples detection results are illustrated in Figure 5. The two shown on the left are accurate detections relative to the ground truth, those on right are mistakes. Comparing the detected parts to the test images' ground-truth annotations, we find that while many of the detections have significant errors (*e.g.* those in Figure 5), many detections are reasonably accurate. As it is pointless to try to classify these false detections, we run the classification on the more accurate detections as described below.

### 6.3. Subordinate Categorization

We now describe our subordinate categorization results. We establish a baseline using a pyramidal histogram of color-SIFT words approach (using the VLFeat toolbox [46] implementation), providing it the ground-truth bounding box to assist in localizing the bird. The performance across test-training splits is 37.12% mean-average precision. Anecdotally, this approach is comparable to the multiple-kernel learning approach used by Branson *et al.* [9] (37.02% on this same subset of categories). Figure 4(a) shows a confusion matrix for the Baseline PHOW/SVM classifier. Next we turn to Figure 4(c), which illustrates the potential performance of the PNAD-RF (Pose-Normalized Appearance Descriptor coupled with the Random Forest classifier) technique. This approach achieves a mean-average precision across the categories of 66.58% by computing the PNAD features on the ground truth ellipsoids.

Our objective then is to evaluate the same PNAD-RF method on the estimated ellipsoids from our real detections. Figure 6 shows the mean classification accuracy for sets of increasing size. The plot shows that, for the most accurate 20% of the detections, the subordinate classification accuracy was above the baseline performance. For the top 10% of detections, accuracy was as much as 10% higher than that of the baseline. In Figure 4(b), the confusion matrix for the most accurate 20% of the detections is shown, a mean-average precision of 40.25%. We believe that the performance coule be even higher if the birdlet training had a larger pool of training examples to draw upon.

## 7. Conclusion

We have presented an approach for subordinate categorization using a pose-normalized appearance representation founded upon a volumetric part model. We model a basic-level category by its constituent parts (a set of volumetric primitives), then leverage the variation in part shape and appearance properties across a taxonomy to provide the additional cues needed for subordinate-level discrimination.

Our model learns to associate raw image patterns (used in detection) with corresponding volumetric part parameters such as location, scale and orientation. These volumetric parameters implicitly define a mapping from the image pixels into a pose-normalized appearance space, removing view and pose dependencies, thus facilitating effective subordinate categorization.

## References


[1] B. Babenko, S. Branson, and S. Belongie. Similarity Metrics for Categorization: From Monolithic to Category Specific. In *ICCV*, 2009.

[2] A. Bar-Hillel and D. Weinshall. Subordinate Class Recognition Using Relational Object Models. In *NIPS*, 2007.

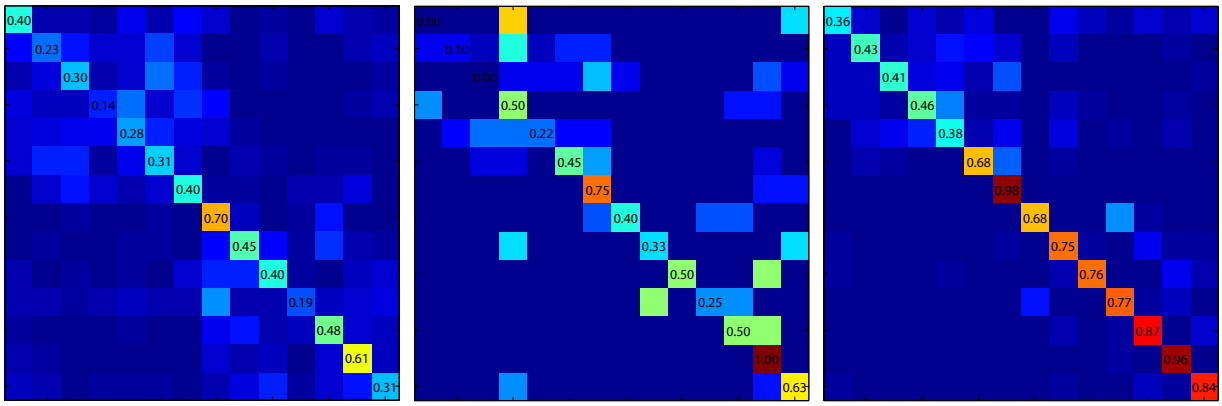
[3] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, 2008.

[4] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, 2010.

[5] I. Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115 – 147, 1987.

[6] I. Biederman, S. Subramaniam, M. Bar, P. Kalocsai, and J. Fiser. Subordinate-level object classification reexamined. *Psychological Research*, 62(2-3), 1999.

[7] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting People Using Mutually Consistent Poselet Activations. In *ECCV*, 2010.




(a) Baseline - PHOW/SVM (b) Detected parts (top 20%) - PNADs (c) Ground truth part locations - PNADs

Figure 4. **Classification Confusion Matrices.** Depicts the classification for the following techniques (a) the PHOW/SVM Baseline (37.12% MAP), (b) the PNAD-RF performance on the top 20% of detections (40.25% MAP), and (c) the PNAD-RF performance on the ground truth part locations (66.58% MAP).


[8] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

[9] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *ECCV*, 2010.

[10] L. Breimano. Random Forests. *Machine Learning*, 45:5–32, 2001.

[11] M. Burl, T. K. Leung, and P. Perona. Face Localization via Shape Statistics. In *Workshop on Automatic Face and Gesture Recognition*, 1995.

[12] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models, 2001.

[14] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial Priors for Part-Based Recognition using Statistical Models. In *CVPR*, 2005.

[15] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[16] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? In *ECCV*, 2010.

[17] 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.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), June 2010.

[19] A. Farhadi, I. Endres, and D. Hoiem. Attribute-Centric Recognition for Cross-category Generalization. In *CVPR*, 2010.

[20] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009.

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

[22] L. Fei-Fei, R. Fergus, and P. Perona. One-Shot learning of object categories. *PAMI*, pages 594–611, 2006.

[23] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *To appear in PAMI*.

[24] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61:55–79, 2005.

[25] A. Ferencz, E. G. Learned-Miller, and J. Malik. Building a Classification Cascade for Visual Identification from One Example. In *ICCV*, 2005.

[26] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic Label Sharing for Learning with Many Categories. In *ECCV*, 2010.

[27] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005.

[28] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 2010.

[29] M. Fischler and R. Elschlager. The Representation and Matching of Pictorial Structures. *Computers, IEEE Transactions on*, C-22(1):67 – 92, Jan 1973.

[30] J. J. Gibson. *The Ecological Approach To Visual Perception*. Houghton Mifflin, Boston, 1979.

[31] G. Griffin and P. Perona. Learning and Using Taxonomies for Fast Visual Categorization. In *CVPR*, 2008.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update, 2009.

[33] M. J. Jones and T. Poggio. Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes. *IJCV*, 29, 1998.

[34] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009.

[35] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.

[36] H.-T. Lin, C.-J. Lin, and R. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68, 2007.

[37] D. Marr and H. K. Nishihara. Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 200(1140):pp. 269–294, 1978.

[38] M. Marszałek and C. Schmid. Constructing Category Hierarchies for Visual Recognition. In *ECCV*, 2008.

[39] G. Martínez-Muñoz, N. Larios, E. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T. Dietterich. Dictionary-Free Categorization of Very Similar Objects via Stacked Evidence Trees. In *CVPR*, 2009.

[40] E. Miller, N. Matsakis, and P. Viola. Learning from One Example Through Shared Densities on Transforms. In *CVPR*, 2000.

[41] M.-E. Nilsback and A. Zisserman. A Visual Vocabulary for Flower Classification. In *CVPR*, 2006.

[42] M.-E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 2008.

[43] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic Objects in Natural Categories. *Cognitive Psychology*, 8(3):382 – 439, 1976.

[44] M. J. Tarr, P. Williams, W. G. Hayward, and I. Gauthier. Three-dimensional Object Recognition is Viewpoint Dependent. *Nature Neuroscience*, 1(4):275–277, 1998.

[45] S. Todorovic and N. Ahuja. Learning Subcategory Relevances for Category Recognition. In *CVPR*, 2008.

[46] A. Vedaldi and B. Fulkerson. VLFeat: An Open and Portable Library of Computer Vision Algorithms. `http://www.vlfeat.org/`, 2008.

[47] G. Wang and D. Forsyth. Joint Learning of Visual Attributes, Object Classes and Visual Saliency. In *ICCV*, 2009.

[48] M. Weber, M. Welling, and P. Perona. Unsupervised Learning of Models for Recognition. In *ECCV*, 2000.

[49] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[50] A. Zweig and D. Weinshall. Exploiting Object Hierarchy: Combining Models from Different Category Levels. In *ICCV*, 2007.