Learning Categorical Shape from Captioned Images

Tom S. H. Lee, Sanja Fidler, Alex Levinshtein, Sven Dickinson
Department of Computer Science
University of Toronto
Toronto, Canada
{tshlee, fidler, babalex, sven}@cs.toronto.edu

***Abstract*—Given a set of captioned images of cluttered scenes containing various objects in different positions and scales, we learn named contour models of object categories without relying on bounding box annotation. We extend a recent language-vision integration framework that finds spatial configurations of image features that co-occur with words in image captions. By substituting appearance features with local contour features, object categories are recognized by a contour model that grows along the object's boundary. Experiments on ETHZ are presented to show that 1) the extended framework is better able to learn named visual categories whose within-class variation is better captured by a shape model than an appearance model; and 2) typical object recognition methods fail when manually annotated bounding boxes are unavailable.**

***Keywords*-Language-Vision integration, Image annotation, Perceptual grouping, Object categorization, Semi-supervised shape learning**

I. Introduction

Learning visual category models from training images is now standard practice in the object categorization community [1], [2]. Such systems typically rely on a strong degree of supervision, including cluttered scenes with labeled bounding boxes placed around objects of interest, or alternatively, scenes in which the labeled object of interest is largely front and centre (allowing the image boundary to serve as an effective bounding box). However, as the scope of the recognition task scales up to many thousands of objects, the burden of manually annotating the large number of required training images becomes prohibitive.

Captioned images are ubiquitous on the web and in certain image collections, and offer a powerful semi-supervised mechanism for learning category models without the need for labeled bounding boxes or image cropping. Unfortunately, any given image-caption pair may be unreliable, providing very weak or even erroneous training data. For example, the caption's nouns may refer to objects that don't appear in the image, while the more salient objects in the image might not even be referred to in the caption. However, across a large training set, recurring correspondences between particular objects appearing in the images and particular nouns appearing in the captions of those same images can be assumed to be salient. Such correspondences can therefore be analyzed to yield visual category models as well as the names of those categories.

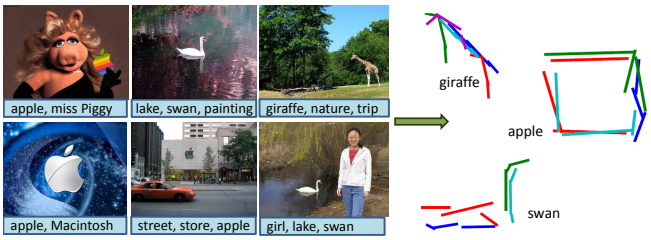


Figure 1. Given a set of captioned images of objects appearing in different positions and scales, we learn named contour models without bounding boxes.

In [3], such a framework was proposed which learned structured visual object models from captioned training data. A learned visual model was captured in a graph, in which nodes represent SIFT features and edges represent spatial relations among the SIFT features, including relative position, orientation, and scale. When applied to captioned image collections, the choice of SIFT to characterize a node ultimately constrained the system to learn the structure and names of exemplars rather than categories. For example, the system learned the logos (and team names) of NHL teams from captioned action images (containing players whose jerseys contained the logos) taken from the NHL website, and learned the models and names of famous buildings and landmarks from around the world from captioned image landmark collections. The question we pose is whether such a framework can be extended to learn named categorical models based on shape rather than appearance.

Extending the framework of Jamieson *et al.* [3] poses a number of significant challenges, including 1) the choice of a suitable structured shape representation that can accommodate within-class deformation, articulation, and occlusion, and 2) coping with the tremendous ambiguity of local shape features relative to appearance-based features such as SIFT. For a categorical shape representation, we adopt and extend the contour representation introduced by Ferrari *et al.* [4], replacing the star graph with a more generic graph with spatial relations between any pair of nodes, and adding multi-scale feature extraction. Like Jamieson *et al.*, we "grow" visual models that repeatedly co-occur with caption words across training images. However, for any pair of spatially related

contour features representing an initial model, there may be many false positives across training images due to their inherent lack of specificity compared to appearance-based features. As a result, we introduce a powerful bottom-up heuristic that can focus search for recurring shape features that are likely to represent the boundaries of objects.

We evaluate our approach head-to-head with Jamieson *et al*., and demonstrate that on a standard benchmark, it clearly outperforms Jamieson *et al*. in terms of learning visual categories in which shape is more invariant than appearance. Since we have based our shape representation on that of Ferrari *et al*., we demonstrate that for the task of learning visual models with correct object labels but without the aid of bounding boxes, our approach outperforms Ferrari *et al*.'s approach, which depends heavily on the strong supervision offered by a bounding box. Finally, we demonstrate the robustness of our approach under image caption noise.

II. Related Work

There is a vast literature on language-vision integration, and the related problems of object category modeling, recognition and localization. While it is beyond the scope of this paper to provide a full review of these topics, we will focus on the two subfields most related to our work: 1) using language or text to discover associations between visual and textual features, and 2) weakly- or semi-supervised object category learning using part-based models in support of image annotation.

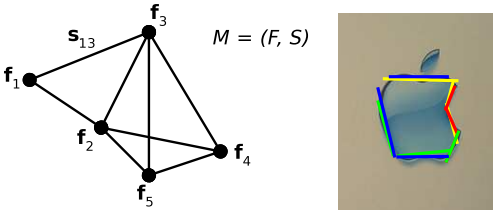
Automatic image annotation systems attempt to discover correlations between words and visual features in a set of image-text pairs, *e.g*., Barnard *et al*. [5] and Duygulu *et al*. [6]. Such systems typically model objects as a mixture of appearance-based features in which common configurations are not captured by explicit spatial relations but rather by co-occurrence statistics, *e.g*., Carneiro *et al*. [7], Monay and Gatica-Perez [8], and Quattoni *et al*. [9]. The relatively high dimensionality of appearance-based features, *e.g*., SIFT, means that image features are relatively unambiguous and therefore explicit relations are often unnecessary. The most similar work to ours is Jamieson *et al*. [3], whose framework used language to recover an explicit graph-based structural appearance model from captions training images. While the structural model captured explicit relations, its reliance on appearance-based local features rendered it far more suitable for learning named exemplars rather than named categories. We attempt to extend that framework to support the learning of visual object categories based on a structural shape representation, which is far more invariant to within-class variation than a structured appearance representation.

Learning a visual category model in the absence of image captions has received considerable attention from the recognition community. Most current approaches assume that bounding boxes around the objects are given [1], [10], [2]. In our domain, we want to avoid such strict supervision, and learn objects from cluttered scenes without any a priori information about location and scale. Moreover, labeling is assumed to be noisy in the sense that a noun in the caption may or may not refer to an object in its associated image, and an object in the image may or may not be referred to in the caption. Given coarse location and scale information, a number of frameworks have learned structured models in terms of parts and relations without requiring part labeling [11], [12], [13]. However, most of these approaches, like Jamieson *et al*. [3], rely on the distinctiveness of local appearance-based features, such as image patches or SIFT, because such features allow the search space to be aggressively pruned, thereby reducing the complexity of the task. Furthermore, [11], [12], [13] constrain object models to a star structure, while in our work object features can be connected using a denser graph whose number of vertices and edge structure are inferred from images without supervision.

There are a few approaches which attempt to learn object models without the strong supervision provided by a bounding box. For example, Todorovic and Ahuja [14] propose a powerful framework for unsupervised modeling based on tree matching using a region-based object representation. Lee and Grauman [15] perform object discovery over images of multiple categories, using matching local appearance patches to anchor an initial set of edge fragments. In our approach, we do not rely on sparse discriminative features, and instead use only dense contour features which capture object categories better. Both Leordeanu *et al*. [16] and Payet and Todorovic [17] use only shape features to learn object models with weak supervision. In [16], object models consisting of hundreds of fully interconnected features require class labels for learning, while in [17], clusters of matching pairs of contour features and spatial relations are found in an unsupervised manner. While our visual representation also consists only of contours, we take an integrated approach where learning is guided by both bottom-up segmentation and image caption text to achieve a comparatively efficient way to initialize visual clusters among multiple categories. The key concept here is that in our approach, we focus on the construction of only those models that are referred to (i.e., named) in the captions, as opposed to mining a much larger space of possible regularities across a set of images, regardless of whether they are salient or not.

III. Overview

Given a set of captioned images, we learn object models that co-occur with words, and use the learned models to detect and annotate objects in uncaptioned images. In Section IV, we describe the object model, a graph in which vertices are local contour features and edges encode pairwise spatial relations between features. Section V describes how object models are detected in an image by matching the



(a) Object model graph (b) Model instance

Figure 2. An object is modeled as a graph M over contour features F with pairwise spatial relations S.

model's contour features to those in the image. In a cluttered image, typically a large number of ambiguous features match individually to model features, creating an intractable search space. Efficient detection is achieved by using the model's spatial relations to prune out unlikely matches.

Section VI describes how we learn object models that co-occur with words. In a graph-growing process, an initial model representing a small part of the object is iteratively grown, primarily along the object's boundary, to cover the object. Features found in the vicinity of existing model matches are added to the model if they recur with spatially consistent relations, thus making the model more object-specific and strengthening its co-occurrence with the given word. When no more consistently recurring features can be found in the vicinity, co-occurrence can no longer be improved and the final model is returned.

Model growth strongly depends on the initial model representing a small part of the object. Whereas in [3] a spatially related pair of appearance-based features was distinctive enough to represent a salient part, a related pair of contour features is relatively ambiguous, *i.e.*, a contour representation of an object part is often very similar to, and thus easily confused with, contours from background clutter or even other objects. To ensure that model growth begins from an object part, we use bottom-up segmentation as a powerful heuristic to focus search on contours that are likely to represent an object boundary.

IV. Object Model

An object model is denoted as $M = (F, S)$, where M is a graph with a vertex set of contour features $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_T\}$ where T is unknown, and an edge set of pairwise relations $S \subseteq \{\mathbf{s}_{ij} : 0 \leq i < j \leq T\}$. Figure 2(a) shows an example of such a structure. A (undirected) spatial relation $\mathbf{s}_{ij}$ between features $\mathbf{f}_i$ and $\mathbf{f}_j$ may exist for any feature pair. To maintain a spatially coherent object description, we require that M be a connected graph. Typically, a dense set of relations is learned for objects having spatially consistent features and relations, resulting in a relatively distinctive model, while more deformable and articulating objects result in a sparser, more flexible model.

Feature extraction and description. Local contour features are extracted from an image using the method of Ferrari *et al*. [18]. Features are scale-invariant descriptions of line-segment abstractions of image contours, where edgel-chains are partitioned into line segments (*e.g.*, Figure 2(b)). Because linear partitioning is scale-dependent, we extract features at multiple scales to obtain a robust bottom-up description of an image. Efficient matching (Section V) and learning (Section VI) is facilitated by a discrete vocabulary of codewords [18], whereby similar contour features are represented by the same codeword. By performing feature comparisons at the codeword level, many feature similarity computations are saved.

Spatial relations. A spatial relation $\mathbf{s}_{ij}$ encodes the distance u_{ij}, relative direction v_{ij}, and relative scale w_{ij} between features $\mathbf{f}_i$ and $\mathbf{f}_j$, *i.e.*, $\mathbf{s}_{ij} = (u_{ij}, v_{ij}, w_{ij})$. Letting $\mathbf{x}_i$ and s_i denote feature position and scale with respect to the image as in [18], respectively, the three components are defined as follows:

$$u_{ij} = \frac{1}{\lambda}||\mathbf{x}_j - \mathbf{x}_i|| \qquad \text{(distance)}$$

$$v_{ij} = \arctan(\mathbf{x}_j - \mathbf{x}_i) \qquad \text{(relative direction)}$$

$$w_{ij} = \frac{1}{\lambda}(s_j - s_i) \qquad \text{(relative scale)}$$

where $\lambda(s_i, s_j)$ normalizes for the feature scales [3].

While features and spatial relations originate from the image, the object model is a prototypical description of the object using these elements, averaging over natural variations present in example images, *e.g.*, due to deformation or viewpoint variation. The following section explains how the model matches to image features under these variations.

V. Detecting Objects

Occurrences of $M = (F, S)$ are detected by matching model features $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_T\}$ to image features subject to spatial relations S. Due to occlusion and feature extraction errors, only a minimum number of model features are required to match. To detect occurrences in complex, cluttered images efficiently, features are matched sequentially, using S to prune out unlikely combinations at each stage.

Each occurrence is associated with a detection score D that measures the confidence of the match, and can be thresholded to obtain a level of precision. Suppose F' is a set of image features which is a potential match to a set of model features F. The detection score with respect to M is defined as a ratio of two quantities (similarly to [12]):

$$D = \frac{p(F'|M)}{p(F'|\text{bg})}. \qquad (1)$$

The numerator is the probability that F' is a true instance of the object (approximated by M), while the denominator is the probability that F' is not an instance of the object. The natural threshold of 1 for the ratio represents a pruning

threshold for the removal of highly unlikely matches. While the quantity $p(F'|\text{bg})$ determines the level of pruning, and is set independently of the object, $\tau \geq 1$ provides a tighter, object-specific threshold for D.

Let a partial matching be denoted using $\mathbf{c}$, a list of T binary indicators, where c_i is 1 exactly when the model feature $\mathbf{f}_i$ is matched. We use $F(\mathbf{c}) \subseteq F$ to indicate the subset of matching model features, and $S(\mathbf{c}) \subseteq S$ to indicate the subset of model relations between any pair of features in $F(\mathbf{c})$. Furthermore, given a set of matching image features F', let S' denote the set of pairwise relations among F'.

We let the object probability $p(F'|M)$ factorize into a feature and spatial component:

$$p(F'|M) = p(F'|F)p(S'|S). \tag{2}$$

Next, we proceed to define the two components. Let $\mathbf{f}'$ denote the image feature matching the model feature $\mathbf{f}$. The feature component $p(F'|F)$ is defined only over $\mathbf{f} \in F(\mathbf{c})$, as follows:

$$p(F'|F) = \prod_{\mathbf{f} \in F(\mathbf{c})} p(\mathbf{f}'|\mathbf{f}), \tag{3}$$

where $p(\mathbf{f}'|\mathbf{f})$ is a Gaussian density over the feature dissimilarity measure $d(\mathbf{f}', \mathbf{f})$ given in [18], with variance $\sigma_f^2 = 2.0$ and zero mean. The dissimilarity $d(\mathbf{f}', \mathbf{f})$ compares two contours using their line segment abstractions, in particular their internal relative positions, orientations, and lengths.

The spatial component $p(S'|S)$ considers only model relations between matched features, *i.e.*, $\mathbf{s}_{ij} \in S(\mathbf{c})$. Given a pair of matching features, let $\mathbf{s}'_{ij}$ denote the spatial relation in the image corresponding to $\mathbf{s}_{ij}$. The spatial component is then defined as:

$$p(S'|S) = \prod_{\mathbf{s}_{ij} \in S(\mathbf{c})} p(\mathbf{s}'_{ij}|\mathbf{s}_{ij}), \tag{4}$$

where $p(\mathbf{s}'_{ij}|\mathbf{s}_{ij})$ factors into its distance, relative direction, and relative scale components:

$$p(\mathbf{s}'_{ij}|\mathbf{s}_{ij}) = p(u'_{ij}|u_{ij})p(v'_{ij}|v_{ij})p(w'_{ij}|w_{ij}), \tag{5}$$

each of which is a Gaussian density with means u_{ij}, v_{ij}, w_{ij} and fixed variances $\sigma_u^2 = 0.35, \sigma_v^2 = 0.3, \sigma_w^2 = 0.8$, respectively. The spatial component accounts for variations in spatial relations, *e.g.*, due to deformation or viewpoint.

The background probability $p(F'|\text{bg})$ represents a pruning threshold and is similarly composed only of the matching components as follows:

$$p(F'|\text{bg}) = \prod_{\mathbf{f} \in F(\mathbf{c})} p(\mathbf{f}'|\text{bg}_f) \prod_{\mathbf{s}_{ij} \in S(\mathbf{c})} p(\mathbf{s}'_{ij}|\text{bg}_s), \tag{6}$$

where $p(\mathbf{f}'|\text{bg}_f)$ and $p(\mathbf{s}'_{ij}|\text{bg}_s)$ are fixed values that in the product offset the ratio D.

Given the above definitions, the detection score is equivalent to:

$$D = \prod_{\mathbf{f} \in F(\mathbf{c})} \frac{p(\mathbf{f}'|\mathbf{f})}{p(\mathbf{f}'|\text{bg}_f)} \prod_{\mathbf{s}_{ij} \in S(\mathbf{c})} \frac{p(\mathbf{s}'_{ij}|\mathbf{s}_{ij})}{p(\mathbf{s}'_{ij}|\text{bg}_s)}. \tag{7}$$

Note that since features and spatial relations must individually pass the pruning threshold given by $p(\mathbf{f}'|\text{bg}_f)$ and $p(\mathbf{s}'_{ij}|\text{bg}_s)$, each factor in Equation 7 is greater than 1, *i.e.*, each matching component accumulates evidence by increasing the detection score. Since a partial match has fewer components, it is penalized with a lower score.

Detection algorithm. As in [3], matching is done efficiently by using S as a constraint to prune out unlikely feature combinations. A set of potentially matching features F' is iteratively grown until all model features are matched, or no more matching features can be found. Failed partial matches are rejected and search resumes with a new initialization. Multiple occurrences in the same image are detected by repeating the search over remaining image features.

VI. Learning objects

The graph-growing algorithm in [3] iteratively adds model features to cover the object (*e.g.*, in Figure 3), making the model increasingly object-specific and simultaneously increasing its co-occurrence with the word. Since word-object co-occurrence is our objective in finding salient object models in captioned images, we measure co-occurrence using the score $C_{M,W}$, defined as follows. Given N captioned training images, occurrences of words and objects are summarized in two vectors of length N:

$$\mathbf{w} = \{w_1, \ldots, w_N\}, w_n \in \{0, 1\}$$

indicating the occurrence of W in each image caption, and

$$\mathbf{m} = \{m_1, \ldots, m_N\}, m_n \in [0, 1]$$

indicating the occurrence of M in each image. While word occurrences are binary, object occurrences have soft scores weighted by their detection scores D. When there are multiple object occurrences in one image, the highest-scoring detection is considered.

While object and word occurrences are obviously correlated, objects may or may not be referred to in the caption, and words in the caption may or may not refer to objects in the image. The co-occurrence score $C_{M,W}$ has a probabilistic formulation in which the occurrences $\mathbf{m}, \mathbf{w}$ are generated from the presence ($c = 1$) or absence ($c = 0$) of a common object, defined similarly to [3] as

$$C_{M,W} = \frac{p(\mathbf{m}, \mathbf{w}|c = 1)}{p(\mathbf{m}, \mathbf{w}|c = 0)}. \tag{8}$$

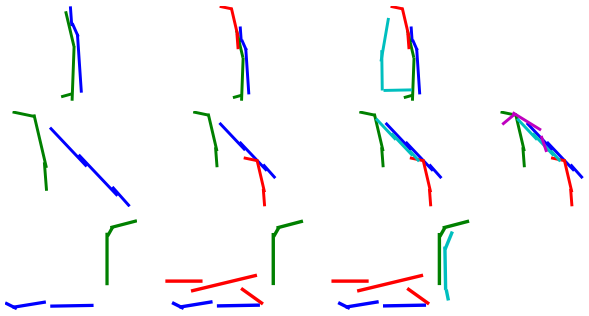


Figure 3. Initial models grow (left to right) to cover the object, increasing in distinctiveness and co-occurrence with words 'bottle', 'giraffe', and 'swan', shown (top to bottom).

The numerator is defined as:

$$p(\mathbf{m}, \mathbf{w}|c=1) = \prod_{n=1}^{N} p(m_n, w_n|c=1) \tag{9}$$

$$= \prod_{n=1}^{N} \sum_{o_n=0,1} p(m_n, w_n|o_n, c=1)p(o_n|c=1)$$

$$= \prod_{n=1}^{N} \sum_{o_n=0,1} p(m_n|o_n, c=1)p(w_n|o_n, c=1)p(o_n|c=1)$$

where o_n is a hidden variable indicating the presence ($o_n = 1$) or absence ($o_n = 0$) of the common object. The quantities $\alpha = p(w|o=1)$ and $\beta = p(w|o=0)$ are the probability that W occurs in the caption when the common object is present or absent, respectively. Similarly $\mu = p(m|o=1)$ and $\nu = p(m|o=0)$ represent the probability of the model M occurring in an image when the common object is present or absent. These parameters control the degree to which $C_{M,W}$ is sensitive to caption noise.

The denominator is defined as:

$$p(\mathbf{m}, \mathbf{w}|c=0) = \prod_{n=1}^{N} p(m_n|c=0)p(w_n|c=0) \tag{10}$$

and acts as a bias term to offset the score $C_{M,W}$.

Graph-growing algorithm. Given an initial model $M^{(0)}$ of two spatially related features representing a small object part, a sequence of successively larger models $M^{(1)}, M^{(2)}, M^{(3)}, \ldots$ is found such that co-occurrence $C_{M^{(k)},W}$ is increasing for $k \geq 0$. More specifically, given occurrences of $M^{(k)}$, a search is performed for a feature in the vicinity that repeats in a consistent spatial relation with respect to the occurrences. Given a candidate shortlist of such features, the candidate with the highest $C_{M^{(k+1)},W}$ is chosen, where $M^{(k+1)}$ is the model with the added candidate feature, provided that $C_{M^{(k+1)},W} > C_{M^{(k)},W}$. When no such candidate exists, the current model is returned as the final model. While the approach described above is greedy, in practice we keep a list of the best few candidates at each iteration to explore in a backtracking fashion.

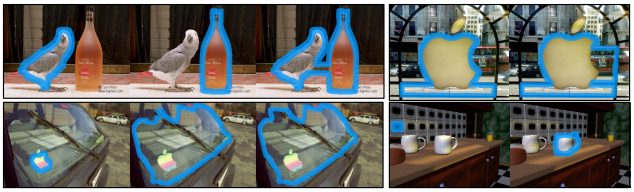


Figure 4. Superpixel Closure [19] returns multiple figure-ground segmentations per image (shown above for 4 images). Model initialization is constrained to contour features that fall along the boundaries of segmentation regions.

Model initialization. It is crucial that $M^{(0)}$ initially represents part of an object so the model can be expanded. The ambiguity of contour features, however, makes it difficult to distinguish salient boundary portions from accidental contours. Bottom-up segmentation offers a powerful heuristic for focusing search over features likely to represent object boundaries. For each image containing the word W, we use Superpixel Closure [19] to extract multiple figure-ground segmentation hypotheses at multiple scales (Figure 4). The boundaries of a figure-ground segmentation hypothesis are used as constraints over features, where only contour features that fall within a small, fixed distance from the region boundary are selected. By initializing $M^{(0)}$ over this subset of features, the heuristic is used to guide the search for promising object parts that can be added to the model.

VII. Evaluation

Following a discussion of the strengths and limitations of our method in Section VII-A, we present a head-to-head comparison of our approach to Jamieson *et al.* [3] in Section VII-B, and two experiments comparing our approach to Ferrari *et al.* [2], on which we have based our shape representation, in Section VII-C. Finally in Section VII-D, we train object models under image caption noise.

Experiments are conducted on the benchmark ETHZ dataset [18], which consists of 255 images of 5 diverse categories labeled with the words 'apple logo', 'bottle', 'giraffe', 'mug', and 'swan'. While bounding boxes are provided with the ETHZ dataset, they are used only for evaluating object localization and *not* for training, unless otherwise noted.

In our evaluation, we have used only the single final model M^* having the highest co-occurrence score $C_{M^*,W}$, in the same manner as in [3]. A possibility is to incorporate combinations of multiple learned models into detection for improved robustness (*e.g.*, models corresponding to different viewpoints), although we have not done this.

A. Qualitative evaluation

In Figure 9 we present sample detections of learned object models for each ETHZ category. Our method is

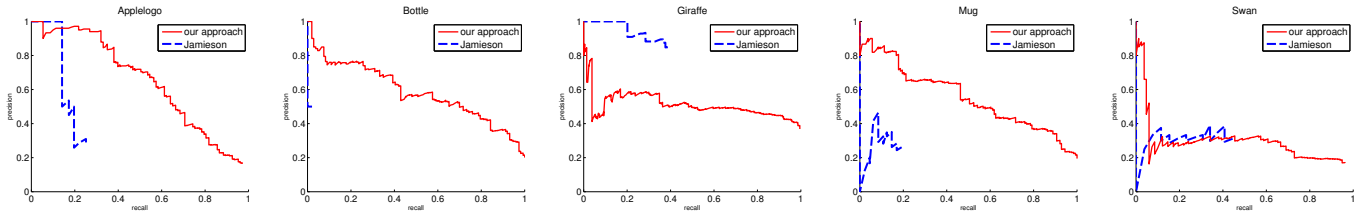


Figure 5. Comparison with Jamieson *et al*. Contour features capture object categories more effectively than appearance-based features.

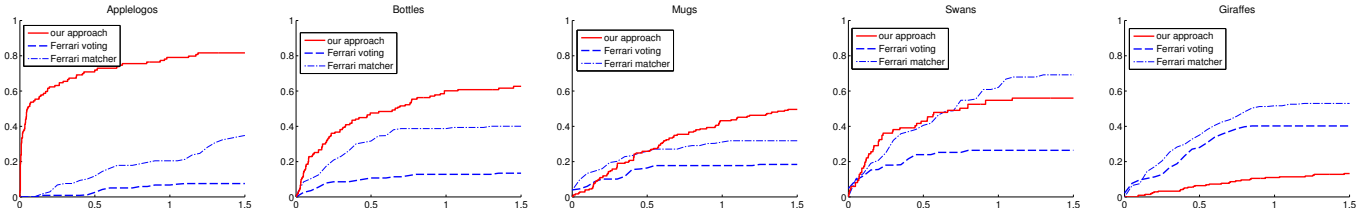


Figure 6. Comparison with Ferrari *et al*., training without bounding boxes.

able to localize objects under large variations in scale and minor changes in viewpoint, orientation, deformation, and articulation. We achieve our best performance with 'apple logo', which features stable contours; our system missed only instances that were severely occluded or rotated by a large amount. 'Bottle' and 'mug' objects have tremendous variation in their surface markings: these objects are similar only in their shape, and our method correctly captures their characteristic contours. One source of false positives, however, is brand labels on bottles, which are easily confused with the bottle boundary due to their proximity. 'Giraffe' and 'swan' pose a significant challenge to our system due to their deformation and articulation. The line-based abstraction of Ferrari *et al*.'s contour features does not always respond in a stable manner to slight changes in curvature, as curve partitioning breakpoints suddenly appear or disappear. When our system encounters highly varying features during the graph-growing process, it tends to terminate growth early to avoid overfitting.

B. *Shape* vs. *appearance models*

A comparison with Jamieson *et al*. [3] allows us to study the effectiveness of shape-based *vs*. appearance-based features for learning named object categories. Our experiments support our hypothesis that shape is more effective. Figure 5 shows the performance of both approaches using precision-recall over correct image annotation, where an annotated image is counted as a true positive if the annotation is correct, *i.e*., a detected occurrence is consistent with the true word label; and counted as a false positive if a detected occurrence is inconsistent with the true label.

Results show that the appearance-based system has difficulty finding recurring object models, especially for 'bottle' and 'mug'. These categories exhibit virtually no regularities in colour, texture, or surface markings. Instead, the appearance-based system found recurring texture on the body of the giraffe (although overfit), and company slogan text appearing in a limited number of 'applelogo' images. Recurring patterns in the water were found for 'swan' images, but these were similar to parallel strokes in leaves and grass, giving rise to poor precision. While appearance may have a limited ability to describe some object categories, contours were found to be more effective overall.

C. *Training without bounding boxes*

Our next experiments compare our approach with that of Ferrari *et al*. [2] under two settings described below. We follow the strict evaluation protocol in [2] (Pascal criterion of 50% intersection-over-union bounding box overlap), and report in Figure 6 detection rate (DR) against false positives per image (FPPI). For interest, we include performance for Ferrari *et al*.'s initial hough voting stage (lower curve) and the final verification stage (higher curve).

In the first setting we train under realistic conditions where manual annotations (bounding boxes) are not available to the system. Results in Figure 6 show that we generally outperform Ferrari *et al*., which is unable to handle training images where objects do not appear in consistent positions and scales, namely, 'apple logo', 'bottle', and 'mug'. It is clear that Ferrari *et al*.'s method strongly depends on the availability of manually annotated bounding boxes. Ferrari *et al*. performs better for 'giraffe' and 'swan', as the respective objects typically occupy most of the image, *i.e*., the image boundary serves as an effective bounding box.

Our second setting examines the scenario where bounding boxes are provided for training. Although our system is not designed to use the information given by bounding boxes, it is interesting to include a discussion of the results. While our method works well for categories with relatively stable contour representations, *e.g*., 'apple logo', the high variation in deformable and articulating categories present a significant challenge to our model-growing algorithm, and

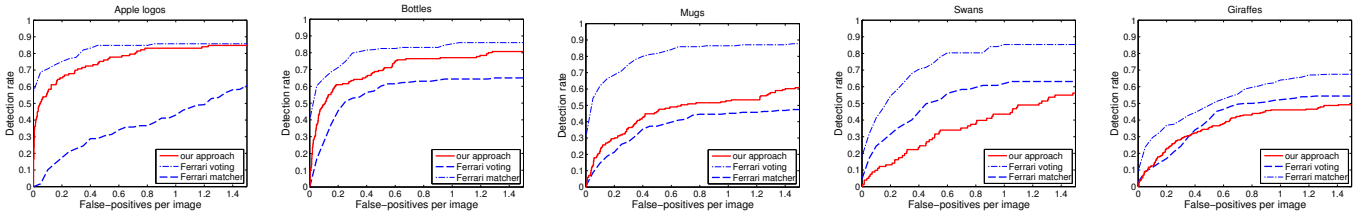


Figure 7. Comparison with Ferrari *et al.*, training with bounding boxes.

our performance is worse than that of Ferrari *et al.*'s (Figure 7). Since models grow incrementally in size, the algorithm is faced with the choice of adding weak, ambiguous contours, and hence may terminate growth early to avoid overfitting, resulting in the strict bounding box overlap criterion not being met. In contrast, Ferrari *et al.*'s Hough-based method has a more global view by having each model feature vote independently for the object centroid. However, this explicitly takes advantage of information from the bounding box, and thus performs an easier task than ours.

D. Image caption noise

In our final experiment, we approximate real-world captions by adding noise to the ETHZ word labels (referred to below as captions). Recall from Section VI that the co-occurrence score assumes that an object is likely to be referenced in the caption with probability α, and an object has a chance of being referenced even when it is absent, with probability β. We corrupt the 5 original words in the ETHZ captions as follows. First, with probability α, words are substituted with a random word (not restricted to the 5 original words). For example, some 'apple logo' images no longer have the word 'apple logo' in the captions. Secondly, with probability β, an original word is appended to captions not originally containing that word. For example, some non-'apple logo' images now have the word 'apple logo' in their captions. Distinct levels of noise are quantified with a percentage p, where $p\%$ reflects the value $\alpha = 1 - \frac{p}{100}$ and the value $\beta = \frac{p}{100}$. We run experiments at different noise levels, where captions are corrupted prior to training. Figure 8 reports localization performance as a function of noise levels, and shows that performance for 'apple logo' and 'bottle' remains stable under significant noise.

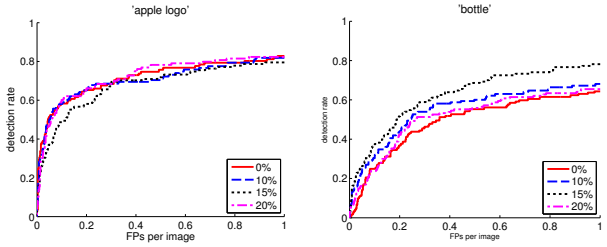


Figure 8. Localization performance of 'apple logo' and 'bottle' as caption noise is increased by corrupting ETHZ word labels. Results show stability under significant noise (up to 20%).

VIII. Conclusion

We have extended the framework of Jamieson *et al.* [3] to learn named categorical models based on shape, and added a focus-of-attention heuristic to cope with the ambiguity of contour features. Using a standard benchmark we have demonstrated that our method is able to handle large variations in scale and minor changes in viewpoint, deformation and articulation. A comparison with Jamieson *et al.* [3] showed that object categories are better captured by shape than appearance. Additionally we outperform methods such as Ferrari *et al.* [2] when training without bounding boxes, and showed that such methods have a strong dependence on manual annotations.

As part of future work, we would like to extend the method in numerous ways to make it applicable to larger datasets. Ultimately, our goal is to learn object class models directly from the web. This entails having a richer representation (shape as well as appearance, more flexibility in the shape features), dealing with multiple classes in a scalable way as well as a more involved text model that could better deal with noisy image tags and surrounding web text.

Acknowledgements. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright notation herein.

References

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models." *PAMI*, vol. 32, no. 9, 2010.

[2] V. Ferrari, F. Jurie, and C. Schmid, "From images to shape models for object detection," *IJCV*, vol. 87, no. 3, pp. 284–303, May 2010.

[3] M. Jamieson, A. Fazly, S. Stevenson, S. Dickinson, and S. Wachsmuth, "Using language to learn structured appear-

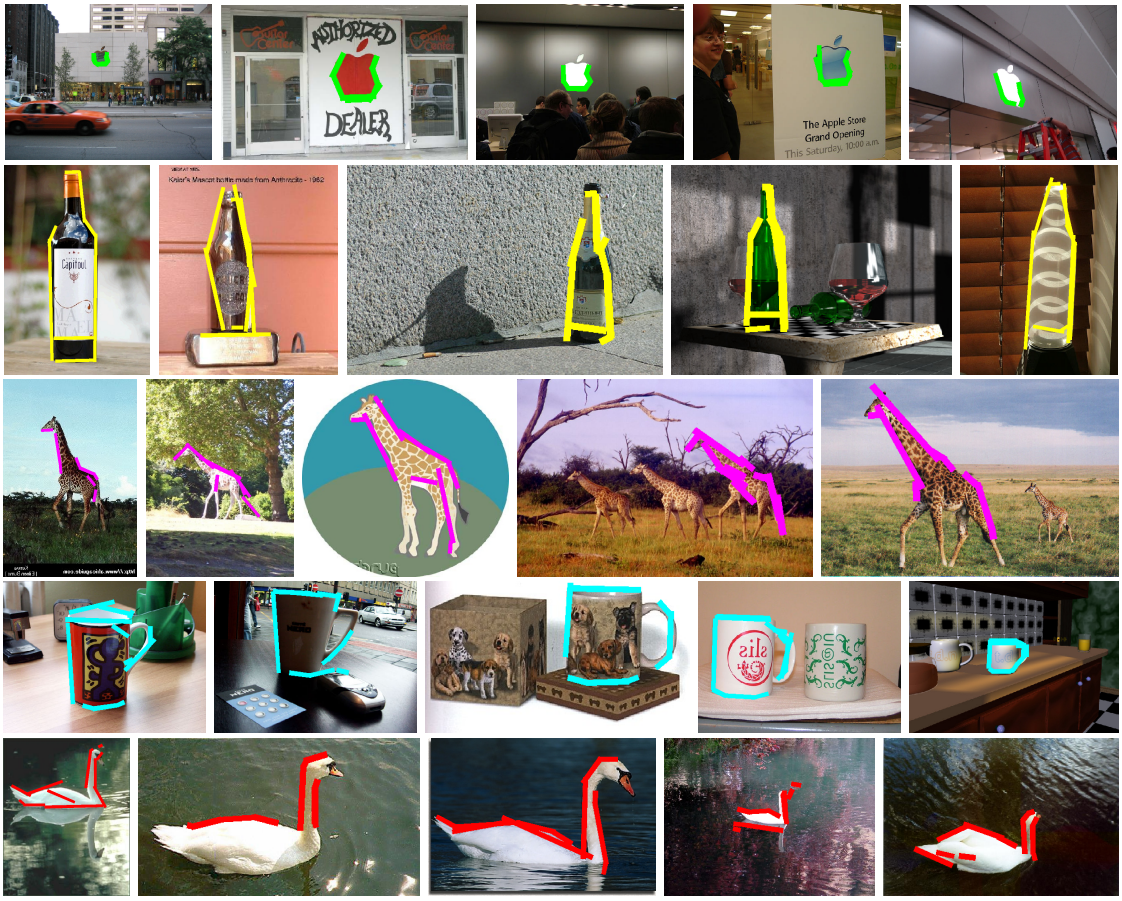


Figure 9. Sample contour detections of 'applelogo', 'bottle', 'giraffe', 'mug', and 'swan' (each in a unique color) in the ETHZ dataset.

ance models for image annotation," *PAMI*, vol. 32, no. 1, pp. 148–164, 2010.

[4] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *PAMI*, pp. 1–16, Nov 2008.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. D. Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *JMLR*, vol. 3, pp. 1107–1135, 2003.

[6] P. Duygulu, K. Barnard, J. D. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *ECCV*, pp. 349–354, 2002.

[7] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *PAMI*, 2007.

[8] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *PAMI*, 2007.

[9] A. Quattoni, M. Collins, and T. Darrell, "Learning visual representations using images with captions," in *CVPR*, 2007.

[10] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *ICCV*, 2007.

[11] D. Crandall and D. Huttenlocher, "Weakly supervised learning of part-based spatial models for visual object recognition," *ECCV*, pp. 16–29, 2006.

[12] R. Fergus, P. Perona, and A. Zisserman, "Weakly supervised scale-invariant learning of models for visual recognition," *IJCV*, vol. 71, no. 3, pp. 273–303, 2007.

[13] S. Bagon, O. Brostovsky, M. Galun, and M. Irani, "Detecting and sketching the common," in *CVPR*, 2010.

[14] S. Todorovic and N. Ahuja, "Unsupervised category modeling, recognition, and segmentation in images." *PAMI*, 2009.

[15] Y. Lee and K. Grauman, "Shape discovery from unlabeled image collections," *CVPR*, 2009.

[16] M. Leordeanu, M. Hebert, and R. Sukthankar, "Beyond local appearance: Category recognition from pairwise interactions of simple features," *CVPR*, 2007.

[17] N. Payet and S. Todorovic, "From a set of shapes to object discovery," *ECCV '10*, 2010.

[18] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Object detection by contour segment networks," *ECCV*, pp. 14–28, 2006.

[19] A. Levinshtein, C. Sminchisescu, and S. Dickinson, "Optimal contour closure by superpixel grouping," *ECCV*, pp. 480–493, 2010.