Attribute-based Vehicle Recognition using Viewpoint-aware Multiple Instance SVMs

Kun Duan
Indiana University
Bloomington, IN USA
kduan@indiana.edu

Luca Marchesotti
Xerox Research Centre Europe
Grenoble, France
luca.marchesotti@xerox.com

David J. Crandall
Indiana University
Bloomington, IN USA
djcran@indiana.edu

Abstract

Vehicle recognition is a challenging task with many useful applications. State-of-the-art methods usually learn discriminative classifiers for different vehicle categories or different viewpoint angles, but little work has explored vehicle recognition using semantic visual attributes. In this paper, we propose a novel iterative multiple instance learning method to model local attributes and viewpoint angles together in the same framework. We expand the standard MI-SVM formulation to incorporate pairwise constraints based on viewpoint relations within positive exemplars. We show that our method is able to generate discriminative and semantic local attributes for vehicle categories. We also show that we can estimate viewpoint labels more accurately than baselines when these annotations are not available in the training set. We test the technique on the Stanford cars and INRIA vehicles datasets, and compare with other methods.

1. Introduction

Several recent papers have studied visual identification of vehicle makes and models [7, 10, 13, 20]. This is an appealing problem because it is a well-defined fine-grained recognition task that has many potential applications. For instance, fine-grained vehicle recognition could automatically annotate images on car sales websites with detailed textual descriptions to allow for better browsing and search, or could identify cars in surveillance video for safety and security applications. But identifying specific vehicle models is difficult, even for humans: most vehicles have similar global shapes and visual cues, so one must rely on subtle differences in local appearance (Figure 1). Another challenge is viewpoint variation: cars are 3D objects whose appearance changes dramatically across different angles. Vehicle recognition also inherits all of the usual complications of fine-grained recognition, including complex and confusing backgrounds, illumination and scale changes, etc.

Recent work has used visual attributes to address fine-grain recognition problems [2,3,19,21]. Attributes are mid-level image representations that are both visually discriminative and semantically meaningful to humans. Attributes are advantageous because they connect low-level visual features with high-level nameable properties, allowing humans to help specify models (*e.g.* in zero-shot learning [15]), and allowing models to produce textual descriptions of objects and scenes [12]. Local attributes, which cue on local image features instead of scene-level features, are particularly relevant for vehicles since they can capture subtle differences between similar categories. A key problem in attribute-based techniques is how to select the attributes themselves. While most work has used attributes suggested by domain efforts, recent work has found that automatically-selected local attributes give better discriminative power [5].

In this paper, we propose to discover local attributes using multiple instance learning and apply our technique to vehicle recognition. A key difference compared to other work is that we explicitly consider viewpoint while selecting attributes. This is motivated by the fact that vehicles are usually photographed from a relatively small set of viewpoints (*e.g.* side, front, back, etc.), and that local attributes are closely connected with viewpoint (*e.g.* a perfectly round-shaped wheel implies a side view.) We assume that viewpoint labels are not necessarily available for training images, so we must hypothesize viewpoints in addition to discovering image regions corresponding to attributes.

To do this, we assume that each training image can be considered as a bag of extracted image regions [5, 21], and our goal is to find the subset of regions that are discriminative. To encourage discovered regions to be semantically meaningful, we augment the classical Multiple Instance SVM (MI-SVM) model [1] with constraints on the positions of image regions across images: for any pair of images within the same vehicle class, we assume that a good attribute should occur at similar places on the vehicle if the viewpoints of the two images are the same, or at similar places after an image transformation if their viewpoints are

(a) Fiat Panda (INRIA dataset)

(b) Ford Escape (Stanford Cars)

Figure 1: Sample images from two categories of *(a)* the INRIA vehicles dataset, and *(b)* the Stanford cars dataset. The INRIA set is significantly more challenging, with cars of the same make and model but different years mixed together.

different. The result of this technique is a collection of localized visual attributes for vehicle images.

To summarize, the contributions of this paper are: 1) to learn discriminative and semantic local attributes for vehicle categories; 2) to devise a multiple instance learning framework with constraints to discover local attributes; 3) to learn the appearance of discrete viewpoints when their annotations are not available; and 4) to show that our learned attributes improve the performance of object recognition.

2. Related Work

Work outside the context of attribute discovery has explored local discriminative regions for image classification. For example, Yao et al [22] use a random forest with dense sampling to discover discriminative regions. The random forest combines thousands of region classifiers together, thus improving classification compared with only low-level image features. In contrast, our approach treats each image as a bag of "regions" and applies multiple instance learning to find the most discriminative ones. Our modification to the MI-SVM model allows pairwise constraints on object geometry, and thus is more likely to find image regions that are both discriminative and semantically meaningful. Maji and Shakhnarovich [17] propose an approach for "part discovery" on landmark images, by collecting pairs of user click annotations. They use exemplar SVMs [18] to find salient regions, while using click pair information to jointly infer object parts. Their method does not optimize classification accuracy, while our approach learns a set of regions by maximizing the classification performance through a multiple instance learning framework.

A number of recent papers on attribute discovery are relevant to our proposed approach, but all have important differences. Gu and Ren [9] learn viewpoint angles and vehicle classifiers at the same time, but they do not consider requiring these models to be semantically-meaningful (at either global or local levels). Duan et al [5] learn local attributes in the context of animal species recognition, but they do not consider multiple viewpoints, and their method relies on multiple features (contour, shape, color, etc.) with carefully learned weights for each feature channel. In contrast, we model local attributes and viewpoint angles together in a single framework, such that the local attribute discovery will help to model the viewpoint angles, and vice versa. Perhaps the most relevant work to ours is that of Sharma et al [19], which automatically mines a collection of parts and corresponding templates for recognizing human attributes and actions. However this method assumes that the attribute labels for training images are given, while we assume only category labels are available, and we want to model local attributes and viewpoint angles at the same time.

3. Approach

We propose a method for automatically discovering discriminative local attributes for vehicle categories. The discovered collection of local attributes serves as a new image representation, which improves vehicle classification performance when fused together with low-level features using the method in [4]. Meanwhile, the discovered attributes can be assigned semantic meanings, allowing novel cross-modal applications such as querying vehicles using textual descriptions. We first describe a technique based on the classic MI-SVM model (Section 3.1), and then we extend it by introducing pairwise constraints (Section 3.2). Finally we describe how to learn the latent viewpoint angles when these annotations are not available (Section 3.3).

3.1. MI-SVMs for attribute discovery

Multiple Instance Learning (MIL) is a form of semi-supervised learning in which training instances are grouped into *bags*. The ground-truth labels of the individual instances are unknown, but each bag has a label that is positive if *at least* one instance in the bag is positive, and negative if *all* its instances are negative. Suppose we have a set of bags $\{x_I\}$. The standard Multiple Instance SVM (MI-SVM) [1] is formulated as an optimization,

$$\min_{\mathbf{w},\xi,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_I \xi_I \qquad (1)$$

$$\text{s.t.} \quad Y_I \cdot \max_i(\mathbf{w}\cdot x_I^i + b) \geq 1 - \xi_I,$$

where $\mathbf{w}$ is a feature weight vector, b is a scalar bias, ξ_I is a slack variable corresponding to training bag x_I, x_I^i is the ith training instance of bag x_I, and Y_I is the ground truth label (+1 or -1) of x_I. Intuitively, this is the classic SVM max-margin framework with an additional (soft) constraint that all instances in the negative bags should be classified as negative, and at least one instance in each positive bag should be classified as positive.

Our goal is to find local image regions across the training set that are discriminative — that occur often in one vehicle category but not in another. We can apply the MI-SVM framework to this problem in the following way. Choose a pair of vehicle categories, calling one positive and one negative. We think of each image as a bag with a positive or negative label depending on its category, and then sample many patches from each image to produce instances for each bag. We then solve equation (2), which produces a weight vector but also implicitly chooses positive instances, and these can be viewed as the set of discriminative regions that we are interested in. We can repeat this process for many pairs of categories to produce a set of candidate attributes.

3.2. MI-SVMs with constraints

A problem with the above approach is that discovered regions may not correspond to the same part of the vehicle, and thus may not have semantic meaning, and also that more than one region may be selected in each positive image. To address these problems, we add constraints to encourage spatial consistency, requiring regions to occur in roughly the same position on the vehicle by adding pairwise spatial constraints among instances in the positive bag. But since viewpoints vary across images, we must explicitly model viewpoint in order to compare spatial positions.

Our model. Let $v_I \in \mathcal{V}$ denote the viewpoint label of image (bag) I, where we assume that $\mathcal{V}$ is a small set of possible discrete viewpoints. For now we assume the viewpoint labels are given; we discuss how to handle unknown viewpoint labels in Section 3.3. We formulate the attribute discovery problem using MI-SVMs, with additional pairwise spatial constraints among positive instances that encourage the spatial consistency property, as illustrated in Figure 2. Suppose that we knew which instance in each positive bag should be part of the attribute, and denote this region x_I^* for bag I. Then we could solve a separate MI-SVM problem for each individual viewpoint $v \in \mathcal{V}$,

$$\min_{\{\mathbf{w}^{(v)}, \xi, b^{(v)}\}} \frac{1}{2}\|\mathbf{w}^{(v)}\|^2 + C^{(v)} \sum_{I \in \mathcal{I}^{(v)}} \xi_I \qquad (2)$$
$$\text{s.t. } \forall I \in \mathcal{I}^{(v)}, \quad Y_I \cdot (\mathbf{w}^{(v)} \cdot x_I^* + b^{(v)}) \geq 1 - \xi_I,$$

where $\mathcal{I}^{(v)}$ is the set of images having viewpoint label v, *i.e.* $\mathcal{I}^{(v)} = \{I | v_I = v\}$.

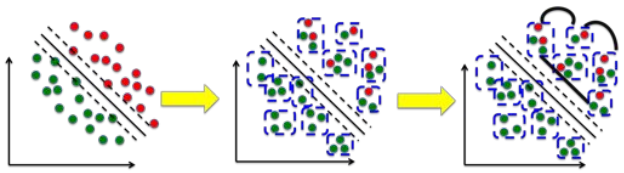


Figure 2: Visualization of SVM models: standard SVM *(left)*, standard MI-SVM *(middle)*, and our MI-SVM with constraints *(right)* between instances in each positive bag. For recognizing vehicles given their viewpoint angles, we define the constraints such that two selected region candidates must come from consistent locations on the vehicles.

Now suppose the weight vectors and biases for each viewpoint were already known, so that we need to estimate the x_I^* for each bag I. We want to do this in a way that encourages spatial consistency. We pose this problem as inference on a Conditional Random Field (CRF) [14]. Let l_I be a scalar variable which takes a value from the region indices in image I. We define an energy function to measure the compatibility of a given assignment of variables to l_I,

$$E(\{l_I\}|\{v_I\}) = \sum_I \phi(l_I|v_I) + \sum_{I,J} \delta(l_I, l_J|v_I, v_J), \qquad (3)$$

where the first set of terms in the summation measures how well the selected regions are modeled by the MI-SVM,

$$\phi(l_I|v_I) = -(\mathbf{w}^{(v_I)} \cdot x_I^{l_I} + b^{(v_I)}),$$

and the pairwise terms encourage positive regions to be at about the same spatial position on the car. If the viewpoint labels between two images are the same, then measuring this distance is a simple matter of comparing image coordinates. If the labels are different, then we need to apply a transformation so that the two coordinate systems are comparable. In particular, our pairwise function is,

$$\delta(l_I, l_J|v_I, v_J) = \begin{cases} \|\mu(l_I) - \mu(l_J)\|^2, & \text{if } v_I = v_J \\ \|H_{v_I}^{v_J}\mu(l_I) - \mu(l_J)\|^2, & \text{if } v_I \neq v_J, \end{cases}$$

where $\mu(l_I)$ denotes the spatial position of region l_I relative to the vehicle center, and $H_{v_I}^{v_J}$ is a homography matrix. We estimate the homography between two viewpoints by extracting SIFT features [16] from the training images having each viewpoint and running RANSAC [8] on feature correspondences. Finally, to estimate the best region x_I^* for each image I, we minimize equation (3) through CRF inference,

$$\{x_I^*\} = \arg\min_{\{l_I\}} E(\{l_I\}|\{v_I\}). \qquad (4)$$

Of course, in our problem we know neither the SVM parameters or the region selections. We thus solve these iteratively, first finding the weights and biases in equation (2) by holding the region variables fixed, and then solve for the

region variables in equation (4) while holding the SVM parameters fixed. The result is a collection of region selections for all positive training images.

Generating regions. We have not yet addressed how to generate the instances within each bag. Although we could randomly sample patches, in practice this creates many irrelevant regions. We thus use an approach similar to [17], applying a pre-trained deformable part-based model car detector [6] on the training images to produce multiple detections with part locations. We then sample from the part detection bounding boxes to generate region candidates. This is faster than the hierarchical segmentation in [5] and produces regions that are more likely to be on the vehicles.

Generating multiple attributes. The above procedure can be used to find the best attribute for a given pair of categories, but in practice we want to generate multiple attribute hypotheses. To do this, we first find the best attribute by solving for $\{x_I^*\}$ using the iterative procedure described above. To find a second attribute, we modify the unary term of equation (3) so that a large constant penalty is paid for selecting an l_I that was chosen as part of the earlier attribute. In our experiments, we repeat this procedure 5 times to produce 5 attribute candidates per pair of categories.

3.3. Recovering Viewpoint Angles

We now consider the case in which the viewpoint labels $\{v_I\}$ are not available ahead of time, so we need to estimate the viewpoint label of each image in addition to the local attributes. We first initialize the viewpoint labels with K-means clustering using global image gradient features (*e.g.* dense SIFT [16]) with $K = |\mathcal{V}|$. Then, after each new attribute is discovered, we update the viewpoint label of each image. To do this, we apply the attribute detectors that have been found so far across all viewpoint angles on the discovered region, choose the best detector, and assign that viewpoint to the region. For all of the discovered regions in an image, we collect all such viewpoint predictions, and use these to vote for the viewpoint of the image.

4. Experiments

We consider two datasets in our experiments: **Stanford cars** [20] with 14 car categories (and 68 training and 34 test images in each category); and **INRIA vehicles** [11] with 29 categories and a total of 10,000 images equally split into training and test sets. There is viewpoint angle bias in both datasets (*e.g.* images in Stanford cars are mostly from $45°$ and $135°$). In Stanford cars, each category consists of car images of the same make, model and year, and bounding box annotations and 8 discrete viewpoint labels are also provided. The INRIA dataset does not have viewpoint labels, and the images in a category are only guaranteed to be the same make and model, not necessarily from the same year.

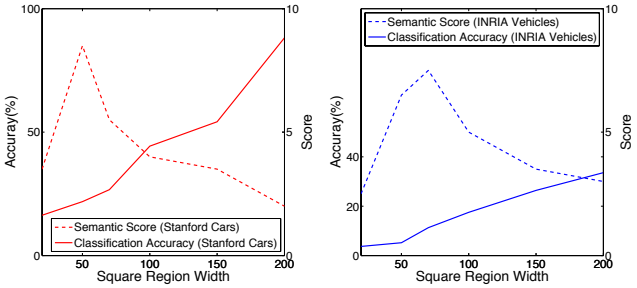


Figure 3: Relationship between region size and *(solid line)* the performance of image classification using single regions, and *(dashed line)* the semantic meaningfulness as judged by humans, for Stanford *(left)* and INRIA *(right)*.

On both datasets, we extract dense SIFT and color histogram features for each region candidate and compute corresponding Fisher vectors using 32D Gaussian mixture models. We also extract these features on the whole image with a three-layer spatial pyramid as a baseline. Note that [5] uses hierarchical segmentation to generate image region candidates, and different types of features (shape, contour, color, gradient, etc) are extracted from the segments. In our case, since the sampled regions are all rectangles, we only use gradient and color features.

4.1. Single attributes

To validate the region pooling parameters and test how our sampling strategy is related to accuracy, we test *single region performance*, where we train multi-class linear SVM classifiers on *all* image region features, using category labels as training labels for classifying vehicle categories. We observe that the performance for classifying single regions decreases as region size decreases (Figure 3). This make intuitive sense because discriminative information is lost when the image is broken into small pieces. For example, it is difficult to tell the difference between two vehicle categories if only parts of the wheels are given.

We also wanted to measure the relationship between region size and whether or not a region is semantically meaningful. To do this, we conducted a simple experiment on Mechanical Turk where image regions of varying sizes were shown, and users were asked to rate (on a scale of 1-10) whether the region corresponded to a meaningful part of the vehicle or not. Results are also shown in Figure 3. We found that semantic meaning suffers if regions are too big or too small: too small cannot capture useful image content, while too big loses interpretability and locality of attributes. Based on these results, we set the region size for the remainder of the experiments in order to maximize the semantic meaningfulness of our image region candidates, generating 50×50 regions for Stanford and 70×70 regions for INRIA.

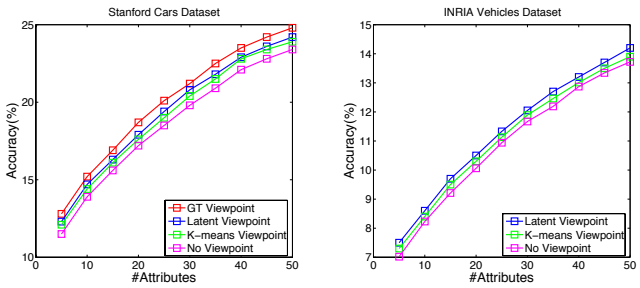


Figure 4: Classification accuracy with different numbers of discovered attributes and different techniques for handling viewpoints, for Stanford *(left)* and INRIA *(right)* datasets.

4.2. Multiple attributes

To use multiple attributes for classification, we aggregate the discovered attributes and use them to build a new representation for each training image. To do this, we apply each attribute classifier on a held-out validation set, and collect all attribute detection scores. We build a $T = (K \times A)$ table, where K is the number of categories and A is the number of attributes. If more than half of the images in a category k have attribute a, then we set $T(k, a) = 1$, otherwise to 0. We use T for nearest neighbor classification.

Our framework can be used to generate multiple attributes by learning MI-SVM on different category pairs and by forcing candidate regions not to overlap (Section 3). However not all such attribute candidates are beneficial to the overall classification performance, so we use an attribute selection method similar to [5] to select the subset of best ones for classification. When a new attribute is generated, we keep it if it improves the overall classification accuracy on a held-out validation set; otherwise it is dropped.

Semantic filtering and naming. We post process these attribute candidates to name them using human feedback from Amazon Mechanical Turk. We present the attribute visualizations generated from each viewpoint to human subjects with their cropped bounding boxes, and put them in a single image gallery if they correspond to the same attribute. Specifically, we asked each subject for the part name, a descriptive word, and their confidence score (on a 1-5 scale) as well. We remove the non-semantic ones if the average confidence score is lower than 3. Every candidate was shown to 5 human users, and the names of the attributes were determined by the majority of the feedbacks.

Category classification results. We studied classification accuracy according to number of detected attributes, as shown in Figure 4. We also compare several attribute selection methods requiring different degrees of viewpoint supervision. **GT Viewpoint** uses the ground truth viewpoint labels in the training set using our technique of Section 3.2. **No Viewpoint** completely ignores viewpoint information in the vehicle discovery process (*i.e.* all images are assumed to have the same viewpoint label). **K-means Viewpoint** runs K-means using global image features to assign initial viewpoint labels without any further update (*i.e.* performs only the initialization phase of Section 3.3). Finally, **Latent Viewpoint** uses our full model, treating viewpoint labels as unknown latent variables and applying the method in Section 3.3. From the figure, we see that incorporating viewpoints into the model helps classification accuracy across any number of attributes. The best results are achieved when viewpoint is available in ground truth, but our technique that can infer viewpoints automatically performs better than either of the simpler baselines. Note that we use 8 viewpoints in these experiments and the INRIA vehicles dataset does not have ground truth viewpoint annotations, so we only report results for the other three methods.

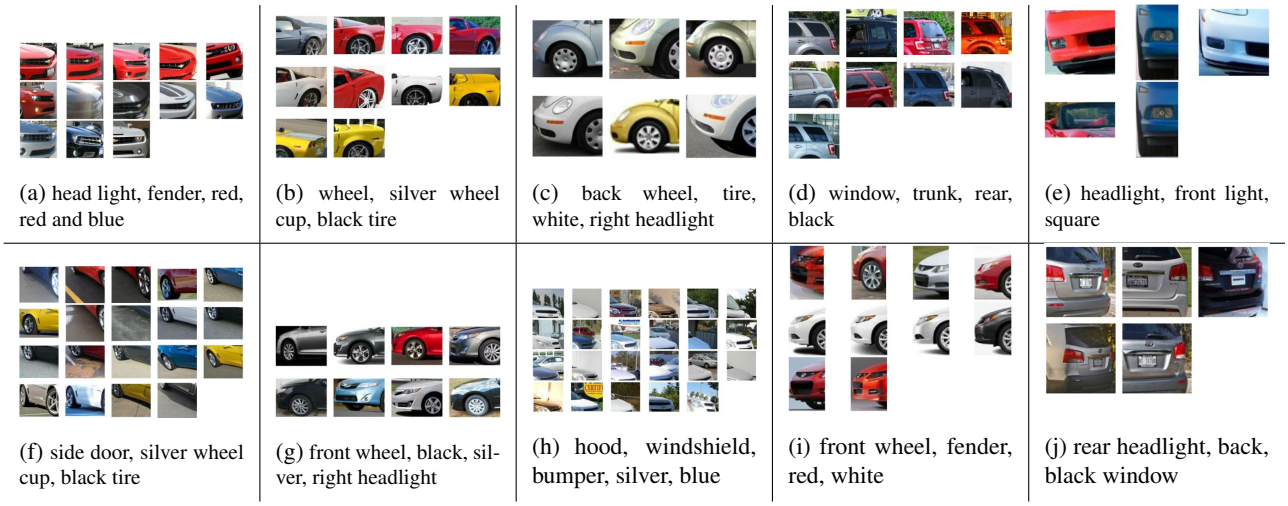
Combining with low level features. We achieve better results by combining the attribute features with low-level Fisher vector features [4]. We use a simple blending scheme on the normalized scores of each test image as $S = \alpha \cdot S_{low} + (1-\alpha) \cdot S_{attr}$, where S_{attr} is the classification score from attributes and S_{low} is the score from the Fisher vectors. We choose the best α using a held-out validation set. On both datasets, we find that combining attributes and low-level features improves classification accuracy compared with just using the low-level features, with an increase from 88.2% to 89.57% on Stanford cars and 33.58% to 34.54% on INRIA, both using 50 attributes. (Note that low-level results on the INRIA set reported in [11] are higher, but they use a much larger mixture model to compute Fisher vectors, so the numbers are not directly comparable.)

Qualitative Results. Figure 5 shows sample local attributes learned using our technique applied on the Stanford cars dataset. These semantic and discriminative visual attributes can be used for automatic image annotation on new images. Figure 6 shows sample tags produced for test images on the Stanford cars dataset.

5. Conclusion

We have presented a novel approach for discovering local visual attributes for vehicle categories and for modeling viewpoint classes at the same time. We have performed systematic experimental evaluations to demonstrate our discovered attributes help to improve baseline classification methods. We showed that our discovered attributes are both discriminative and semantically meaningful, leveraging user feedback on the machine-generated attribute candidates. In future work, we will explore more useful applications of local attributes (*e.g.* image retrieval, automatically caption generation, etc.) and will study incorporating local attributes into vehicle detectors.

Acknowledgements. This work was supported in part by



(a) head light, fender, red, red and blue

(b) wheel, silver wheel cup, black tire

(c) back wheel, tire, white, right headlight

(d) window, trunk, rear, black

(e) headlight, front light, square

(f) side door, silver wheel cup, black tire

(g) front wheel, black, silver, right headlight

(h) hood, windshield, bumper, silver, blue

(i) front wheel, fender, red, white

(j) rear headlight, back, black window

Figure 5: Examples of automatically generated local attributes for the Stanford cars dataset. Each panel represents one discovered local attribute for a particular viewpoint of the vehicle category, with names coming from Mechanical Turk users.

Figure 6: Examples of vehicle annotation results on new images.

the National Science Foundation (IIS-1253549) and by the IU Office of the Vice Provost for Research through the Faculty Research Support Program. The Xerox CVG group is supported in part by the ANR Fire-ID project. Part of this work was done while Kun Duan was an intern at XRCE.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002. 1, 2

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*. Springer, 2010. 1

[3] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*. Springer, 2010. 1

[4] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *CVPR*, 2011. 2, 5

[5] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *CVPR*, 2012. 1, 2, 4, 5

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 4

[7] R. Feris, J. Petterson, B. Siddiquie, L. Brown, and S. Pankanti. Large-scale vehicle detection in challenging urban surveillance environments. In *WACV*, 2011. 1

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981. 3

[9] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010. 2

[10] D. M. Jang and M. Turk. Car-rec: A real time car recognition system. In *WACV*, 2011. 1

[11] J. Krapac, F. Perronnin, T. Furon, and H. Jégou. Instance classication with prototype selection. In *ICMR*, 2014. 4, 5

[12] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. BabyTalk: Understanding and generating simple image descriptions. *PAMI*, 2013. 1

[13] C.-H. Kuo and R. Nevatia. Robust multi-view car detection using unsupervised sub-categorization. In *WACV*, 2009. 1

[14] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 3

[15] C. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. *PAMI*, 2013. 1

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3, 4

[17] S. Maji and G. Shakhnarovich. Part discovery from partial correspondence. In *CVPR*, 2013. 2, 4

[18] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2

[19] G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *CVPR*, 2013. 1, 2

[20] M. Stark, J. Krause, B. Pepik, D. Meger, J. J. Little, B. Schiele, and D. Koller. Fine-grained categorization for 3d scene understanding. In *BMVC*, September 2012. 1, 4

[21] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 1

[22] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 2