Object Instance Sharing by Enhanced Bounding Box Correspondence

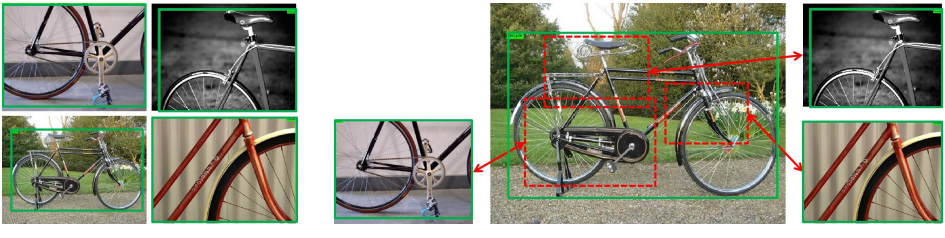
Santosh K. Divvala
santosh@cs.cmu.edu
Alexei A. Efros
efros@cs.cmu.edu
Martial Hebert
hebert@ri.cmu.edu

The Robotics Institute,
Carnegie Mellon University
Pittsburgh, PA. USA.

Abstract

Most contemporary object detection approaches assume each object instance in the training data to be uniquely represented by a single bounding box. In this paper, we go beyond this conventional view by allowing an object instance to be described by multiple bounding boxes. The new bounding box annotations are determined based on the alignment of an object instance with the other training instances in the dataset. Our proposal enables the training data to be reused multiple times for training richer multi-component category models. We operationalize this idea by two complementary operations: bounding box shrinking, which finds subregions of an object instance that could be shared; and bounding box enlarging, which enlarges object instances to include local contextual cues. We empirically validate our approach on the PASCAL VOC detection dataset.

1 Introduction



(a) 4 "bicycles", no correspondence (b) 4 "bicycles" with correspondence

Figure 1: Four images taken directly from the PASCAL VOC dataset with human-annotated bounding boxes (green). (a) Despite each of the four instances having the same label ("bicycle"), their bounding boxes are not aligned so they cannot be used together as training data for a single classifier. (b) By bringing the instances into correspondence (red boxes), we can now use them to provide more training data to a classifier. This is especially important for heavily occluded/truncated instances, where data shortage is a big problem.

Consider the task of building a sliding-window object detector. The standard learning-based approach is to first turn each human-labeled bounding box into a feature vector using some feature descriptor, e.g. HOG, and then train a classifier, e.g. SVM, on a stack of these

 http://dx.doi.org/10.5244/C.26.60

Figure 2: (left) A bicycle instance with its ground-truth bounding box shown in solid green. (center) Four (of the 25) subcategories discovered by our approach (few sample instances within each subcategory are shown). We allow the bicycle instance to be used multiple times with different bounding box representation for training the subcategory models. The different bounding box extents used per subcategory model are color coded accordingly e.g., subcategory3's match is shown using red dotted box, subcategory4's match shown in red dashed box, etc. (right) Subcategory1 shown after adaptively enlarging the bounding box to include local contextual cues around it.

feature vectors to discriminate them from the rest of the visual world. This is a reasonable strategy for older datasets, such as "INRIA person", where object instances are largely in correspondence, i.e. aligned such that each feature vector dimension has the same visual meaning for all object instances. However, modern datasets, such as PASCAL VOC [21], are much less restricted and do not guarantee good correspondence, with often huge variations between annotated bounding box instances, as can be seen on Figure 1(a).

The way modern approaches usually tackle this problem is by using mixture models [2, 8, 15, 30]. The idea is to somehow segregate instances within a category into disjoint groups (subcategories) and then train separate classifiers for each such subcategory. Each subcategory has reduced appearance diversity (via improved alignment), leading to a simpler learning problem. The recent success of the discriminatively-trained mixture model framework of Felzenszwalb et al., [8] has led to the wide popularity of such models for object detection [14, 17, 18, 20, 23]. Applying such model to the four images in Figure 1(a) would likely result in each being assigned to a separate subcategory and trained with others of its kind. While reasonable, this assumes that a lot of training data is available for each subcategory. But this is often not the case, especially for occluded/truncated instances (to paraphrase Tolstoy: all good instance look the same, each occluded instance is occluded in a different way).

What we propose in this paper is the idea of *training data reuse*. Conceptually, we would like to allow different object subcategories to be able to share (subregions of) each others training instances by providing *extra* correspondences between instances that were not part of the original human-supplied bounding box annotations, as shown in Figure 1(b). We operationalize this by two complementary operations: bounding box shrinking, which aims to find subregions of an instance that could be shared; and bounding box enlarging, which aims to create new subcategories by enlarging instances to include their local context. We show that these operations create more training data for each subcategory, and thus improve object detection performance, especially for occluded/truncated instances.

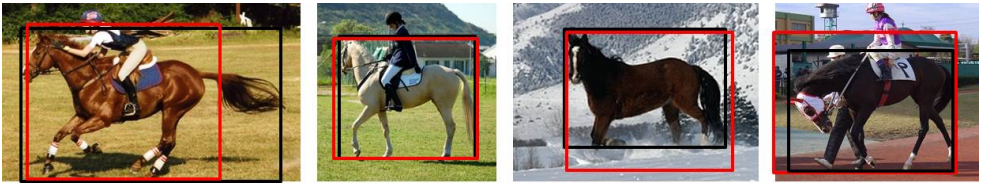


Figure 3: Given four instances from the profile-view horse subcategory (black bounding box indicates ground-truth), latent bounding box fitting method [8, 27] searches for the box representation (red box) that best aligns with the rest of the instances. In this case, the tail of the first horse instance is ignored, while the missing feet of the third horse are hallucinated by using an extended box.

1.1 Overview

Consider the image shown in Figure 2(left). The human-labeled "bicycle" bounding box is indicated by the solid green box. Given this ground-truth framing for the object instance, it is most similar to instances in the "45°-view bicycle" subcategory, so, in a standard mixture-model detector, it would be assigned to subcategory1. However, by relaxing the bounding box framing for this instance, subregions of it can also match to the other subcategory models (subcategory2, subcategory3, subcategory4) as shown using the red bounding boxes. Furthermore, looking *outside* the bounding box might also allow us to capture consistencies in the local context surrounding the object, discovering new subcategories such as "person riding a bicycle" (subcategory5). This observation suggests that by relaxing the human bounding box annotation, one can allow each training instance to be reused multiple times, with different bounding box extents. The new bounding boxes can either be cropped or enlarged versions of the original annotation depending on the alignment with the other training instances.

Of course, the idea of treating human-labeled bounding box annotations as something less than "ground-truth" is not new in object detection. In fact the very criterion for evaluating detection performance in PASCAL VOC allows for just 50% overlap between the predicted detection and the ground-truth bounding box to account for poor alignment due to inaccuracies and arbitrariness of human annotations [7]. In [8, 27], improvement in detection was demonstrated by using latent bounding box fitting, where the human-annotated bounding box is treated as being partially *latent* i.e., the bounding box is allowed to move within a local neighborhood (down to 70% overlap). Intuitively, this can be understood as locally "wiggling" the bounding box representation such that it best *aligns* with the rest of the object instances within a category (or subcategory) as shown in Figure 3.

In this paper, we apply a very similar mechanism, but rather than just making local adjustments, we use it to *search* for bounding box representations that capture new correspondences between instances in the training data. The main difference is that the latent bounding box fitting assumes that each object instance is represented by a *single* bounding box belonging to a single subcategory, whereas our aim is to find *many different* bounding boxes for the same instance, so that it can be shared across multiple subcategories.

The idea of reusing object instances is particularly attractive in gathering extra data for subcategories composed of truncated instances. Truncation is a common occurrence in modern datasets where the object of interest lies partially outside the image area or is occluded e.g., the bicycle instances in Figure 1. Analogous to the heavy-tailed distribution of object categories [24], most training instances within a category are in canonical viewpoints and poses. Due to this lack of sufficient training samples, most detectors do not perform well on truncated instances. Our approach allows canonical non-truncated instances to be reused,

providing extra data in training the subcategories corresponding to truncated instances. As shown in Figure 2, the impoverished bicycle handle subcategory can now use the bicycle handle from the (more common) frontal-view bicycle instance (see Figure 5 for more examples).

In the special case where the new bounding box is larger than the object instance, the extra spatial extent will capture information about the local context of that object. Context plays an important role in aiding object detection [5]. Information around the bounding box often provides useful contextual cues (local pixel context [3, 29]). Nonetheless most sliding-window detection approaches continue to use features computed only within the object bounding box to train the classifier. This is because the local context around the bounding box is highly multimodal for the harder PASCAL or MIT-SUN09 datasets e.g., a horse jumping over a fence appears in a different context compared to a close-up horse shot. However, this could be overcome by simply using a large number of subcategories, as we will show in this paper.

1.2 Related Work

The simplest way of reusing instances is by perturbing existing instances [8, 12, 16, 22] e.g., creating shifted, rotated, or mirrored versions. Most related is the recent work on instance-sharing [10, 18, 19, 28], with the key difference that our focus is on reusing instances within the **same** category and dataset. Also related are the works on transfer learning, where the idea is to reuse data via sharing model parameters or features [11, 24, 31].

There have been a few recent works addressing truncation. Girshick et al., [13] proposed an extension of the deformable parts model of Felzenszwalb et al., [8]. However, their approach involves a hand-defined grammar specific for 'person' class. The training procedure described in [26] involves additional supervision as it requires manually extending the bounding boxes to indicate how far each truncated box ought to extend into the boundary region (and thus presents results only for two of the 20 VOC classes).

There has been a renewed interest towards incorporating local contextual cues [17, 23, 25] for training object models. In [23], detectors are trained for visual phrases that are composed of objects and the typical context surrounding them e.g., "person riding horse", "dog lying on sofa", etc. However their proposed method requires human-annotation of the phrases. In [17], adaptive local context models around the object of interest are separately learned and subsequently used to post-process detection results. We focus on integrating the contextual information directly into the detector, rather than post-processing the detection results.

2 Approach

We begin with the latent SVM based mixture model framework introduced in [8] and then describe details specific to our approach. Consider a classification problem with a training dataset of n labeled examples $D = (< x_1, y_1 >, \ldots, < x_n, y_n >)$, with $y_i \in \{-1, 1\}$. We would like the examples to be clustered into K disjoint subcategories, and separate classifiers to be trained per subcategory. The subcategory memberships z_i are treated as latent variables. The

formulation in [8] uses the following objective function:

$$\arg\min_{w} \frac{1}{2}\sum_{k=1}^{K}||w_k||^2 + C\sum_{i=1}^{n}\varepsilon_i, \tag{1}$$

$$y_i.s_{i,z_i} > 1-\varepsilon_i,\ \varepsilon_i > 0, \tag{2}$$

$$z_i = \arg\max_{k} s_{i,k}, \tag{3}$$

$$s_{i,k} = w_k.\phi_k(x_i) + b_k, \tag{4}$$

where w_k denotes the separating hyperplane for the kth subcategory, $\phi_k(x_i)$ corresponds to the feature representation of an instance i in subcategory k, s_{i,z_i} indicates the score for instance i corresponding to the z_ith subcategory, and C controls the relative weight of the hinge-loss term. Equation (3) is referred to as the latent (discriminative) reclustering step where the latent cluster assignments z_i are iteratively estimated given the model parameters w_k learned in the previous iteration.

Calibration. The output scores $s_{i,k}$ are assumed to be calibrated appropriately for computing the arg max in (3). [4] introduced a calibration step into the original LSVM formulation (1) as:

$$\arg\min_{w,A,B} \frac{1}{2}\sum_{k=1}^{K}||w_k||^2 + C_1\sum_{i=1}^{n}\varepsilon_i + C_2\sum_{k=1}^{K}L_k, \tag{5}$$

$$z_i = \arg\max_{k} g_{i,k}, \quad g_{i,k} = \frac{1}{1+exp(A_k.s_{i,k}+B_k)}, \tag{6}$$

$$L_k = \sum_{i=1}^{n} t_i \log g_{i,k} + (1-t_i)\log(1-g_{i,k}), \quad t_i = Or(W_{i,k}, W_i), \tag{7}$$

where $g_{i,k}$ is the calibrated version of the raw SVM score $s_{i,k}$, $[A_k, B_k]$ are the sigmoid parameters, L_k is the logistic loss function for estimating the sigmoid parameters and $Or(.,.)$ denotes the overlap score. An alternating minimization approach is used for solving it. Given the latent assignment of object instances to subcategories z_i, detectors w_k are first trained for each subcategory. Fixing the detectors w_k, and the latent assignments z_i, the sigmoid parameters A_k, B_k are learned. Finally the detector w_k and the sigmoid A_k, B_k parameters are fixed to update the latent assignments z_i.

2.1 Shrinking Ground-truth Boxes

Our key insight is that it is possible to modify the latent reclustering step in a simple way so as to generate additional samples from a single training instance. The reclustering step involves (i) sliding the K subcategory detectors trained in the previous iteration on the image containing the human-annotated bounding box i and (ii) picking the highest-scoring detection window for each subcategory $W_{i,k}$ (with score $s_{i,k}$) that has at least $T\%$ overlap with the ground-truth window W_i i.e., $Or(W_{i,k}, W_i) > T$ where $Or(W_1, W_2) = \frac{|W_1 \cap W_2|}{|W_1 \cup W_2|} \in [0,1]$ denotes the overlap score [1]. Our key modification to the reclustering step is that the formulation in (1) [8] keeps only *one* of the K detection windows $W_{i,k}$, namely the box W_{i,z_i} with the highest score across all subcategories, while we keep *all* the K windows as long as they pass the overlap test. With this modification, we potentially generate up to K new training samples from each training instance, each of them being aligned to one of the subcategories. We

set T to a low value (10%[1], instead of 70% used in [8]) to encourage valid detections that may have low overlap over an untruncated instance. For example, in the case of the instance shown in Figure 2, the red dotted bounding box corresponding to the bicycle handle will be a valid detection for the bicycle handle subcategory model (subcategory3).

In order to use the multiple samples generated per training instance at each iteration (instead of a single one), we introduce a soft indicator vector $\beta_i = [\beta_{i,1}, \ldots, \beta_{i,k}, \ldots, \beta_{i,K}]$ of length $K \times 1$ into the optimization problem defined in (1). The value of $\beta_{i,k}$ represents the contribution of instance i towards subcategory k and is constrained to range between 0 and 1, with 0 indicating no contribution, and 1 indicating full contribution towards updating the subcategory model w_k. This is formulated as the following objective function:

$$\arg\min_{w,\beta} \frac{1}{2}\sum_{k=1}^{K}||w_k||^2 + C_1\sum_{i=1}^{n}\sum_{k=1}^{K}\beta_{i,k}\varepsilon_{i,k} + C_2\sum_{i=1}^{n}||1-\beta_i||, \tag{8}$$

$$y_i.s_{i,k} > 1-\varepsilon_{i,k},\ \varepsilon_{i,k} > 0, \tag{9}$$

$$s_{i,k} = w_k.\phi_k(x_i) + b_k, \tag{10}$$

$$0 \leq \beta_{i,k} \leq 1. \tag{11}$$

$\beta_{i,k}$ are initialized using the solution of the previous LSVM optimization problem from (1): $\beta_{i,k} = 1$ if $k = z_i$ or $g_{i,k}$ otherwise i.e., all instances originally assigned to the subcategories with their ground-truth bounding box representation will have their contributions set to 1 since they will be fully used, while the samples obtained by describing an instance with new (contracted) boxes will have their β between 0 and 1. Since $g_{i,k}$ is the calibrated SVM score, its value always lies between 0 and 1. The last term in (8) is a regularizer over the indicator vector, which encourages each instance to be reused across multiple subcategories i.e., high regularizer signifies β_i set close to unity.

Solving the optimization problem in (8) for w and β jointly is a non-convex problem. We use an iterative algorithm based on the fact that solving for β given w and for w given β are convex problems. Note that setting the β's to zero for the new samples (those obtained by relaxing the human-annotation) in the above optimization problem simply returns the original LSVM solution.

Implementation details. For solving (8) in our experiments, we iterate only once, as it is sufficient to generate new instances once. Also for improving computation time, we threshold each β so that it will either be 0 or 1. We empirically observed that it is possible for an instance to be reused multiple times with the same bounding box extent, e.g., in the case of the bicycle category, the profile left-view as well as the right-view subcategories confidently score profile view bicycle instances (either left or right facing) with the same bounding box extent. As a result, they both *drift* towards each other. In order to avoid this drift, we non-max suppress[2] the top detections across subcategories that have high overlap with each other. In order to avoid detection windows W_i^k that stride too far outside the ground-truth W_i, we supress detection windows that have high *non-overlap* score with the ground-truth $NOr(W_{i,k}, W_i) = \frac{W_{i,k} - |W_{i,k} \cap W_i|}{W_i}$.

[1] We empirically observed that typical truncations cover at least 10% of an unoccluded fully visible object.

[2] Given multiple overlapping detections, they are sorted by their score and the highest scoring detection is greedily selected while skipping those that have at least 50% overlap with a previously selected detection.

2.2 Enlarging Ground-truth Boxes

As the object instances within each subcategory are tightly aligned in the appearance space, the local regions around them would also be aligned. Therefore we determine the extent of the local region to grow the box adaptively based on the image statistics containing the subcategory instances. Given the object instances within a subcategory, we determine the largest extent $\lambda = [\lambda_{x_1}, \lambda_{y_1}, \lambda_{x_2}, \lambda_{y_2}]$ to which the human-annotated bounding box can be extended (on all four sides) such that the enlarged box is contained entirely within at least 80% of the images in the subcategory. This is done by computing the distance to the image boundary along each side and picking the largest value not exceeding the extent in at least 80% of the instances. All the bounding boxes within the subcategory are grown by this margin:

$$x_1' = x_1 - \lambda_{x_1} W, \quad x_2' = x_2 + \lambda_{x_2} W \quad y_1' = y_1 - \lambda_{y_1} H, \quad y_2' = y_2 + \lambda_{y_2} H, \tag{12}$$

where $W = x_2 - x_1$ and $H = y_2 - y_1$. Figure 6 displays some of the subcategories with their extended bounding boxes. We use the extended bounding boxes as training instances for learning the subcategory models as described in Eq (1). The model dimensions for each subcategory are also extended by a similar margin as in Eq (12) to account for the bounding box extension. We initialize the model using the solution of the previous (unextended bounding box) LSVM optimization function.[3] We emphasize that the latent refitting step during the reclustering process (Eq 3) again plays a crucial role in fixing any misalignment of the extended boxes derived from the initialization step (12) i.e., individual boxes can be adjusted so as to improve alignment with the rest of the instances within the subcategory.

At testing time, we use the extended subcategory models for detecting objects in the conventional sliding-window paradigm. However, prior to evaluation, we shrink the candidate detection windows so as to comply with the evaluation protocol of having at least 50% overlap with the human-annotated (ground-truth) bounding box.

2.3 Initialization

A key step for the success of a mixture model approach is to generate a good initialization of the subcategories. Previous approaches have considered different strategies for initializing subcategories: while some have used extra ground-truth annotations e.g., viewpoint [2, 14], others have relied upon heuristics e.g., aspect-ratio [8]. In [4], it was observed that the common insight shared amongst the different methods is to partition the data such that instances that are visually similar are clustered together. Based on this observation, appearance-based clustering was directly used for initializing the subcategories.

We follow the approach in [4], where all the positive instances within a category are warped to a canonical size for extracting HOG features of fixed dimension, and then unsupervised clustering in this feature space is performed to initialize the subcategories.

3 Experimental Analysis

We evaluated the performance of our approach on the PASCAL VOC 2007 dataset [6]. We used the standard PASCAL VOC comp3 test protocol, which measures detection performance by average precision (AP) over different recall levels. Our experiments are based on

[3] The central region of the extended model is initialized using the model from the previous step and the extended regions are initialized with zeros.

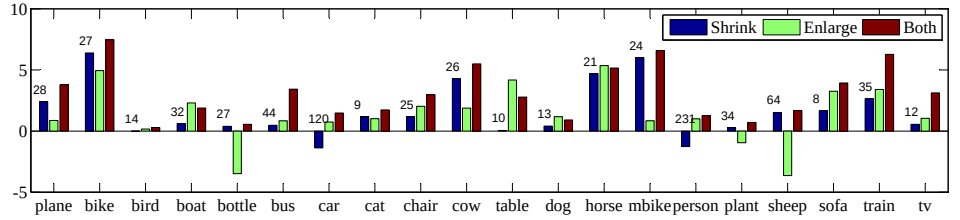


Figure 4: Performance improvement offered by our approach over the baseline (x-axis: 20 VOC classes, y-axis: difference in A.P.). The numbers on top of the blue bar show the percentage increase in the number of data samples (generated via relaxing the human annotation).

Figure 5: Subcategories composed of only a few instances, specifically in case of truncation, can gather more data from other training examples. Each row displays (left) a sample training instance from a subcategory, (right) new samples generated from existing training instances. Red box is the new sample, green box is the human annotation.

the state-of-the-art detector of Felzenszwalb et al., [9]. As our baseline system, we use the detector initialized using appearance-based clustering as used in [4] (with $K = 25$ subcategories). This baseline system with appearance-based clustering used in [4] performs better in comparison to the aspect-ratio based clustering used in [9] (relative improvement of 10.5% across the 20 classes).

Figure 4 compares the results obtained using our approach with respect to the baseline for the 20 PASCAL object categories. The first two bars show the improvements achieved by the shrinking and the enlarging steps respectively. The mean relative improvement (over the baseline) across 20 classes for shrinking is 6.8%, while for enlarging is 5.6%. We also evaluated the result obtained by combining the final subcategory models from each and evaluating them together. The third bar displays the improvements offered by this combined system. The shrinking and enlarging ideas are complimentary to each other and combining them together offers additional boost in performance (mean relative improvement of 12.8%).

Effect of shrinking human annotation. As observed in Figure 4, our shrinking approach almost always improves the results of the baseline system, except for a marginal drop in the

Figure 6: Human-annotated bounding boxes (green box) are automatically enlarged (red box) to leverage local contextual cues (adapted to the subcategory). There is a wide variation in the types of context captured per subcategory. (left) row1: rail tracks for train, row2: wall for sofa, row3: horizontal fence and vertical side bars for horse, row4: sidewalk for bus. (right) row1: people seated at dining table, row2: grass and sky for airplane, row3: person riding bicycle, row4: dining table around a chair. Notice that the local cues do not necessarily correspond to other annotated objects and could include unlabeled regions e.g., rail tracks for train.

case of the car and person category. Atop the blue bar, we show for each class the percentage increase in the number of samples used for training the object model. On average (across the 20 classes), there is a 40% increase in the number of samples used. Figure 5 displays some of the qualitative results for a few impoverished subcategories. We also analyzed the performance gain specifically for detecting truncated instances. We measured the change in A.P. by exclusively evaluating the detector on the truncated instances before and after using the additional samples. We noticed a 30% relative improvement in the mean A.P. across 20 classes.

Effect of enlarging human annotation. As observed in Figure 4, our adaptive enlarging scheme improves performance for all classes except bottle, plant and sheep. Bottles and plants are objects that can typically appear in varied contexts and thus the local context around them can be misleading [5]. Figure 6 displays some qualitative results for a few subcategories. Observe that different subcategories capture different types of local context. For e.g., in case of horse jumping over a fence, the fence and the vertical bars act as discriminative cues in improving the detection of that subcategory. This context would not be valid for a close-up horse face shot subcategory. Thus a monolithic category-based detector would not be able to benefit from local context by simply enlarging the bounding box.

4 Conclusion

Current detection approaches assume each human-labeled bounding box to uniquely describe an object instance. In this paper, we have used the human-labeled bounding box as only a rough indication of object presence. We described each object instance using multiple bounding boxes based on its alignment with other instances in the dataset. Our approach helped in enriching impoverished subcategories with additional data as well as in the inclusion of local contextual cues. In our current implementation, we pool detections across multiple subcategories using a simple sum-pooling based non-max suppression scheme. We

plan to explore learning the spatial relations between the different subcategories for improving the pooling step.

Acknowledgments: This work is supported by NSF Grant IIS0905402.

References

[1] Bogdan Alexe, Viviana Petrescu, and Vittorio Ferrari. Exploiting spatial overlap to efficiently compute appearance distances between image windows. In *NIPS*, 2011.

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

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[4] Santosh K. Divvala, Alexei A. Efros, and Martial Hebert. How important are 'deformable parts' in the deformable parts model? In *ECCV Workshop on Parts and Attributes*, 2012. arXiv:1206.3714.

[5] Santosh Kumar Divvala, Derek Hoiem, James Hays, Alexei A. Efros, and Martial Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge. http://pascallin.ecs.soton.ac.uk/challenges/VOC.

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

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, September 2010.

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

[10] R. Fergus, H. Bernal, Y.Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, 2010.

[11] Tianshi Gao and Daphne Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[12] D.M. Gavrila and J Giebel. Virtual sample generation for template-based shape matching. In *CVPR*, 2001.

[13] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *NIPS*, 2011.

[14] Chunhui Gu and Xiaofeng Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[15] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. Adaptive mixture of local experts. In *Neural Computation*, 1991.

[16] I. Laptev. Improvements of object detection using boosted histograms. In *BMVC*, 2006.

[17] C. Li, D. Parikh, and T. Chen. Extracting adaptive contextual cues from unlabeled regions. In *ICCV*, 2011.

[18] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. In *NIPS*, 2011.

[19] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[20] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010.

[21] PASCAL. The pascal object recognition database collection. Website, 2005. http://www.pascal-network.org/challenges/VOC/.

[22] D. Pomerleau. Neural network perception for mobile robot guidance. In *PhD thesis, Carnegie Mellon University*, 1992.

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

[24] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011.

[25] J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. The visual extent of an object. In *IJCV*, 2012.

[26] A. Vedaldi and A. Zisserman. Structured output regression for detection with partial occulsion. In *NIPS*, 2009.

[27] P. A. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2005.

[28] Gang Wang, David Forsyth, and Derek Hoiem. Comparative object similarity for improved recognition with few or no examples. In *CVPR*, 2010.

[29] Lior Wolf and Stan Bileschi. A critical view of context. *IJCV*, 2006.

[30] Weilong Yang and George Toderici. Discriminative tag learning on youtube videos with latent sub-tags. In *CVPR*, 2011.

[31] Alon Zweig and Daphna Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007.