Object Detection Using Strongly-Supervised Deformable Part Models

Hossein Azizpour[1] and Ivan Laptev[2]

azizpour@kth.se,ivan.laptev@inria.fr
[1]Computer Vision and Active Perception Laboratory (CVAP), KTH, Sweden
[2]INRIA, WILLOW, Laboratoire d'Informatique de l'Ecole Normale Superieure

Abstract. Deformable part-based models [1, 2] achieve state-of-the-art performance for object detection, but rely on heuristic initialization during training due to the optimization of non-convex cost function. This paper investigates limitations of such an initialization and extends earlier methods using additional supervision. We explore strong supervision in terms of annotated object parts and use it to (i) improve model initialization, (ii) optimize model structure, and (iii) handle partial occlusions. Our method is able to deal with sub-optimal and incomplete annotations of object parts and is shown to benefit from semi-supervised learning setups where part-level annotation is provided for a fraction of positive examples only. Experimental results are reported for the detection of six animal classes in PASCAL VOC 2007 and 2010 datasets. We demonstrate significant improvements in detection performance compared to the LSVM [1] and the Poselet [3] object detectors.

1 Introduction

Visual object recognition has achieved substantial progress in recent years, scaling up to thousands of object categories [4] and demonstrating industry-level performance in many applications such as face detection and face recognition. While large-scale visual categorization presents new challenges to the field, many single object classes remain to be very difficult to recognize. For example, the detection of *birds*, *dogs*, *chairs* and *tables* still achieves only modest performance in the recent PASCAL VOC evaluations [5]. This indicates the need of better models and learning methods able to handle objects with large variations in appearance due to intra-class variability, non-rigid deformations, wide range of views and other factors.

Pictorial structures [6] provide a powerful framework for representing objects by non-rigid constellations of parts. Such models have been recently applied to a number of computer vision problems including object recognition [1, 7, 2], human pose estimation [8–10], action recognition [11] and facial feature detection [12]. In particular, discriminatively-trained deformable part-based models (DPMs) [1, 2] have shown excellent performance for object detection tasks. Learning such models, however, involves the optimization of a non-convex cost function over a set of latent variables standing for image locations of object parts and mixture

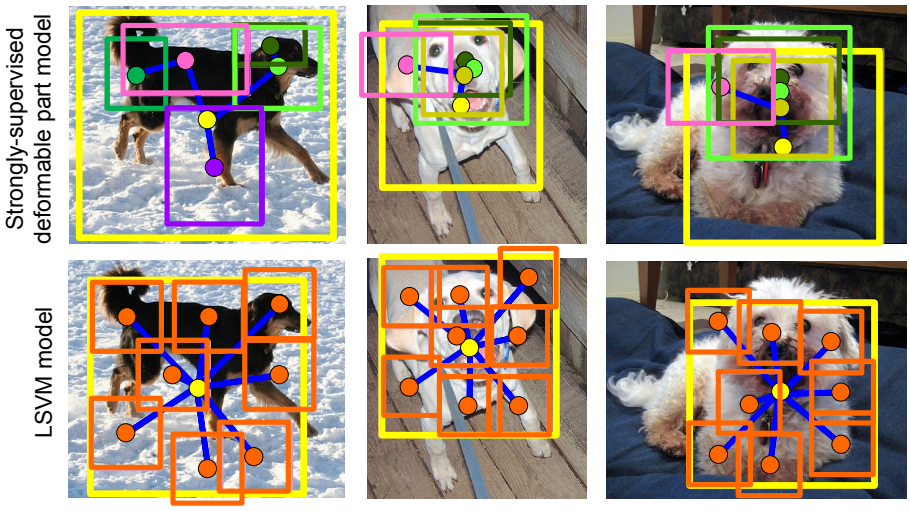

Fig. 1: Dog detection (yellow rectangles) using deformable part-based models. Top: dog detections using the proposed strongly-supervised DPM method. Bottom: dog detections using regular DPMs [1]. Note how the part structure of our model is adapted to the different dog images. In contrast, the original model attempts to explain all three images using the same deformable model.

component assignment. The success of training DPMs, hence, depends on the good initialization of model parts in general [13].

While the latent SVM approach in [1] applies heuristics to automatically initialize model parts, several recent methods explore *strong* part-level supervision. For example, poselet-based methods [3, 14, 15] learn appearance of human body parts from manually annotated limb locations. The joint learning of appearance and deformation parameters in DPMs using part-level supervision has been used for body pose estimation in [10] and object part localization in [16].

The goal of this work is to develop and evaluate a strongly-supervised DPM framework for object detection. Our extensions of existing methods are motivated by the following. First, we expect the use of additional part-level supervision to enhance the performance of current DPM detectors [1, 2]. Second, additional supervision should enable us to construct better, class-specific object models compared to the generic star models (see Figure 1). Finally, compared to the task of part localization [16], object detection does not require precise part recovery and should benefit from a different learning objective compared to [16].

Summarizing our contributions, we use strong supervision to improve the initialization of the LSVM object detector [1] (mixture component associations and part locations) and propose a class-specific optimization of the model structure. We formulate a learning objective which can deal with sub-optimal and incomplete object parts annotations, and which explicitly handles partial object occlusions. We provide extensive evaluation and favorably compare our method

to previous object detectors [1, 3] on the task of detecting six animal classes in PASCAL VOC 2007 and 2010 datasets.

The rest of the paper is organized as follows. Section 2 discusses related work. Next, we describe DPM and its proposed extensions in Section 3. Experimental evaluation of the method is presented in Section 4. Section 5 concludes the paper.

2 Related work

Object detection is a dynamic research area. Originating from work on person detection [17], the HOG descriptors with linear SVM classifiers have been a standard building block for object localization. While HOG represents objects by a rigid template, LSVM detector [1] extends [17] to a discriminatively-trained model of deformable HOG parts. Our work builds on DPMs in [1] and extends it by introducing part-level supervision and optimization of the model structure.

Bag-of-Features (BOF) models have been highly successful for object retrieval [18] and image classification [19]. Application of BOF to object localization, however, has been more problematic due to efficiency reasons. [20] combines BOF with HOG features in the latent SVM DPM framework. Region-level cues and image segmentation have also been explored to improve detection for textured objects such as cats and dogs in [21]. Our work is complementary to these methods in that we use a single feature type only (HOG) while aiming to enrich the structure of the model and to explore part-level supervision.

Strong part-level supervision has been explored in a number of recent works. The work on Poselets [22] uses human limb annotation to learn independent HOG-SVM body-part detectors for person localization. In contrast to this work, we jointly learn part appearance and model deformation parameters and demonstrate improvements over [22] in Section 4. Part-level supervision has been used for human pose estimation [10, 15, 23, 24] and object part localization [16]. While [15, 23] learn part appearance independently, our approach is more related to [10, 16, 24] using strong supervision to train DPMs. The task of part localization addressed in [10, 16, 24], however, is different from the task of object detection in this paper. In particular, the learning objectives in [16, 24] enforce annotation-consistent part localization, which is not directly relevant for object detection and can hurt detection performance if some annotated parts are not discriminative or if their annotation is imprecise. On the contrary, our method optimizes the detection objective and learns discriminative locations and relative weights of parts. Finally, we use mixture of components (part trees) as opposed to the part mixtures in Yang and Ramanan [10]. The use of component mixtures enables us to adapt the structure of the model to different views of the object, as also explored in [25].

In contrast with previous discriminatively-trained DPMs, we (i) explore class-specific optimization of the model structure, (ii) explicitly model part occlusions, and (iii) handle imprecise part-level annotation. We also explore and show benefits of semi-supervised learning where part annotation is provided for a fraction of positive training examples only.

3 Discriminative part-based model

We follow the framework of deformable part models [1, 15, 23, 10] and describe an object by a non-rigid constellation of parts appearance and location. Contrary to the above models we introduce a binary part visibility term in order to explicitly model occlusion. Each part in our model is defined by the location of a bounding box $p_i = (p_i^{x_1}, p_i^{y_1}, p_i^{x_2}, p_i^{y_2})$ in the image and the binary visibility state v_i. One mixture component of the model has a tree structure with nodes U and edges E corresponding to object parts and relations among parts respectively. The score of a model β in the image $\mathbf{I}$ given model parts locations $P = (p_0, ..., p_n)$, and visibility states $V = (v_1, ..., v_n), v_i \in \{0, 1\}$ is defined by the graph energy

$$S(\mathbf{I}, P, V) = \max_{c \in \{1..C\}} S(\mathbf{I}, P, V, \beta_c) \tag{1}$$

$$S(\mathbf{I}, P, V, \beta_i) = \sum_{i \in U} S_A(\mathbf{I}, p_i, v_i, \beta_i) + \sum_{(i,j) \in E} S_D(p_i, p_j, \beta_i)) \tag{2}$$

where the unary term S_A provides appearance score using image features $\phi(\mathbf{I}, p_i)$ (we use HOG features [17])

$$S_A(\mathbf{I}, p_i, v_i, \beta_i) = v_i(F_i.\phi(\mathbf{I}, p_i) + b_i) + (1 - v_i)(F_i^o.\phi(\mathbf{I}, p_i) + b_i^o)$$

and the binary term S_D defines a quadratic deformation cost

$$S_D(p_i, p_j, \beta_i) = d_i.\psi(p_i - p_j)$$

with $\psi(p_i - p_j) = \{dx; dy; dx^2; dy^2\}$ where $dx = p_i^{x_1} - (p_j^{x_1} + \mu_{ij}^x)$ and $dy = p_i^{y_1} - (p_j^{y_1} + \mu_{ij}^y)$. Notably, the score function (1) linearly depends on the model parameters $\beta_c = \{F_0; ...; F_n;\ F_0^o; ...; F_n^o; d_1; ...; d_n; B\}$. To represent multiple appearances of an object, our full model combines a mixture of C trees described by parameters $\beta = \{\beta_1; ...; \beta_C\}$.

Object parts are frequently occluded due to the presence of other objects and self-occlusions. Since occlusions often do not happen at random, the locations of occluded parts may have consistent appearance. We model occlusions by learning separate appearance parameters F^o for occluded parts. The bias terms b_i and b_i^o control the balance between occluded and non-occluded appearance terms in S_A.

3.1 Learning

Given a set of labelled training samples $D = (< x_1, y_1 >, ..., < x_N, y_N >)$, where $x = \{\mathbf{I}, P_x, V_x\}$ is the part-annotated sample ($V_x = 0$ for negative samples) and $y \in \{-1, 1\}$ is class labels, we aim to learn linear parameters of the model β in a discriminative fashion. Towards this goal, similar to [1, 10, 14], we minimize the objective function

$$L_D(\beta) = \frac{1}{2}||\beta||^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i f_\beta(x_i)). \tag{3}$$

We make use of part-level supervision and constrain model parts to be approximately co-located with the manual part annotation (where available) on positive training images. This is achieved by maximizing the scoring function (1) over a subset of part locations and visibility states Z_p, Z_v consistent with the ground truth part annotations P_x, V_x:

$$f_\beta(x) = \max_{P \in Z_p(x), V \in Z_v(x)} S(\mathbf{I}_x, P, V) \tag{4}$$

$$z_p(x) = \begin{cases} \{p \in \mathbb{P} | O(p, p_x) > t_{ovp}\} & \text{if } p_x \text{ available} \\ \mathbb{P} & \text{otherwise} \end{cases} \qquad z_v(x) = \begin{cases} 1 & v_x = 1 \\ \{0, 1\} & v_x = 0 \end{cases}$$

where $\mathbb{P}$ is the set of all possible locations for a part bounding box. The overlap between two bounding boxes $O(p_i, p_j)$ is defined as the intersection over union of their areas, $t_{ovp} = 0.3$ defines a fixed overlap threshold. Note that when annotation is not available for a part, $v_x = 0$, we enforce no constraint. This can help us cope with missing part annotations due to annotators' errors. Note that, as opposed to [16, 24], we have used a loss function which only penalizes a low score for the whole object, while [16, 24] penalizes any instances with different part locations to the annotation. The latter can be helpful for the task of pose estimation while our choice of loss functions allows for down-weighting a part which is not descriptive. This property appears important given the potentially noisy annotations and possibly non-discriminative appearance of some annotated parts.

Before optimizing for the objective function (3), we cluster our positive samples based on their pose, assign a structure to each mixture component and initialize part filters (see details in Sections 3.2 and 3.3). The objective in (3) is known to be non-convex due to the latent variables at positive samples [1]. We solve the optimization problem adopting a coordinate-descent approach in which the latent variables for positive samples are fixed at the first step so that the formulation becomes convex in β. Then, in the second step, we apply stochastic gradient descent (SGD) to estimate model parameters. The steps of the algorithm can be seen in Algorithm 1. In practice we have used a modified version the Latent SVM implementation in [1] and trained models by harvesting hard negative samples from negative training images.

3.2 Pose clustering

Mixture models (components) enable modelling of intra-class variation in the appearance of different samples. These variations are due to strong view-point changes, class subcategories and non-rigid deformations. Assignment of positive samples to components in LSVM formulation is a non-convex optimization and thus sensitive to initialization. [1] proposes simple heuristic and initializes assignments by grouping positive samples based on the aspect ratio of their bounding boxes. This strategy, however, is suboptimal as illustrated in Figure 1 where example images of dogs differ substantially in terms of appearance but do have bounding boxes with similar aspect ratio.

Algorithm 1 Learning framework

Require: $D = (< x_1, y_1 >, ..., < x_N, y_N >)$ the training samples and annotations, C number of clusters

-Pose Clustering: Cluster the positive samples to C clusters based on their poses. (section 3.2)

-Model Structure: **for each** cluster c construct a relations graph E_c using an uncertainty measure u (section 3.3). **endfor**

-Filter Initialization: **for each** part p in each cluster (c, E_c) initialize HOG-SVM part F_p and occlusion F_p^o filters. **endfor**

-Coordinate Descent: do

Positive Latent Fixation: Fix all positive latent variables using equation (4)

Harvest hard negatives using the approach in [1]

SGD: Minimize the objective (3) using stochastic gradient descent.

while (hinge loss value in (3) does not change.**)**

In this work we intend to use part annotations to define better assignment of training samples to model components based on the pose. This allows us to align similar parts better within each component which is an important factor when using linear classifiers. Using annotated parts we parametrize the pose θ_x of sample x by the following vector,

$$\theta_x = (p'_1, ..., p'_n, s_1, ..., s_n, a_0, ..., a_n, v_1, ..., v_n). \tag{5}$$

Here $p'_i = \{p_i^{x_1} - p_0^{x_1}; p_i^{y_1} - p_0^{y_1}\}$ is the relative position of ith part w.r.t. the object bounding box (indexed 0), s_i is the relative scale of ith part width to object width, a is the aspect ratio of parts (including object) bounding boxes, and finally v_i is the binary visibility annotation of part i. We take all positive pose vectors and cluster them using a modified k-means clustering algorithm. Since samples may have occluded parts (without any annotation for bounding box) in each maximization step of the algorithm we replace the missing blocks of vector θ by corresponding blocks of their associated cluster center. Under minor assumptions it will preserve the log-likelihood improvement guarantee of EM algorithm and thus its convergence to local minima [26]. To control the effect of each parameter during clustering, we define a weight for each block of θ by a prefixed parameter (W). In addition, since in the mixture model, we are interested in using horizontally mirrored samples for training the same filters (using mirrored descriptors), we need them to fall into the same cluster. For this purpose we modify the distance between each sample x to a cluster c with center μ_c to be $d(x, c) = min(||\theta_x - \mu_c||, ||\theta'_x - \mu_c||)$ where $\theta'_x = (p''_1, ..., p''_n, s_1, ..., s_n, a_0, ..., a_n, v_1, ..., v_n)$. Note that the only dimensions modified in θ' is the relative position $p''_i = \{p_0^{x_2} - p_i^{x_1}; p_0^{y_1} - p_i^{y_1}\}$ of the horizontally mirrored sample.

This is closely related to the work of Johnson et al.[23]. Contrary to [23], (a) we handle the missing data in a single modified k-means while they decouple the clustering of the fully annotated samples from the partially annotated ones; (b) we assign the same pose in mirrored images to the single cluster giving us more

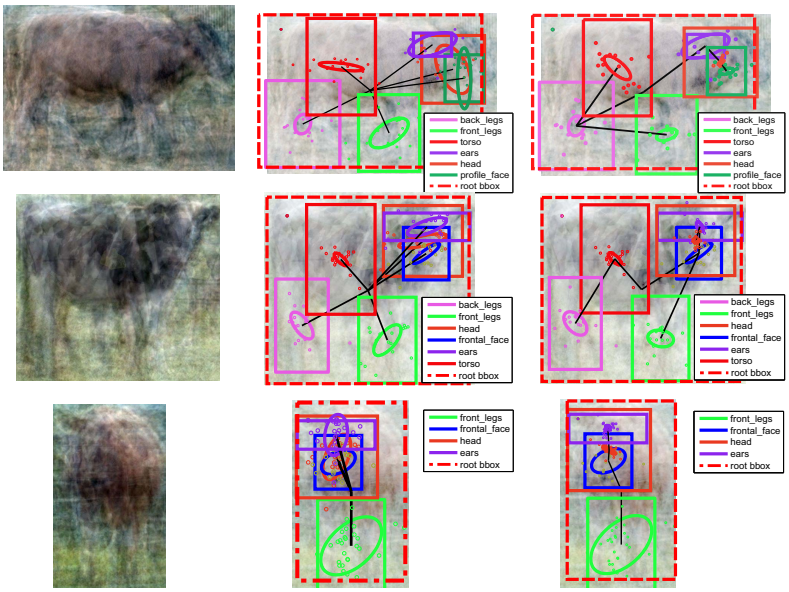

Fig. 2: Illustration of model initialization. Column one: average images of cows from three model components. Column two: star model. Column three: MST-based tree model. Uncertainty in part localization is illustrated by ellipses and is lower for the MST-based tree model compared to the star model. Each row corresponds to one mixture component.

descriptive power; (c) we make use of bounding box annotation by considering aspect ratio in pose vector (potentially putting different part types in different clusters).

Average images generated from our clustering results are illustrated in the first column of Figure 2. The relatively sharp average images indicate consistency among training samples within a component, which can give better discriminative power from more aligned part positions. Note that using part visibility as part of the pose vector helps us to assign systematically similar occluded samples (e.g. cat upper body) to a separate cluster.

3.3 Model structure

The structure of the part based models (dependency graphs) is usually set manually [7, 1, 14, 20, 24, 10], either to a star model [7, 1, 14], hierarchy of coarse to fine relations [20], or, in the case of human pose estimation, to a graph capturing the skeletal configuration of body parts [24, 10]. However, in that case one should define a tree for each class-component which might not be intuitive or optimal. Here we propose to design an optimization procedure for constructing a dependency graph using object part annotations.
As described in [7] the statistically optimal part relations in a *generative* pictorial

structure model can be achieved by optimizing the connections over the following prior probability of samples (assuming each sample is generated independently),

$$E^* = \arg\max_E \prod_{(u_i,u_j)\in E} c(u_i,u_j) = \arg\min_E \sum_{(u_i,u_j)\in E} -log\ c(u_i,u_j) \tag{6}$$

$$c(u_i,u_j) = \prod_{k=1}^{N_p} prob(p^k_{u_i}, p^k_{u_j}|d) \tag{7}$$

Where $-log\ c(u_i,u_j)$ measures the uncertainty of a link over samples given the deformation model d, (i.e. how well the part locations are aligned along a graph edge). However, since we optimize our deformation model discriminatively we can not have d beforehand. Since the connectivities in our model are used to capture the spatial configuration, we approximate the above uncertainty value along each possible edge by variance of relative position of the two end parts. Then we construct a fully connected graph $H = (U_H, E_H)$ with $n+1$ nodes corresponding to object and parts bounding boxes, each edge e is defined as a 3-tuple (i,j,w). For each pair of nodes (i,j) we calculate the diagonal covariance matrix C of the translation (dx,dy) between the two corresponding bounding box centers. Inspired by the Harris corner detector [27] we let the weight of each edge to be $w_e = trace^2(C_e) - k\,det(C_e)$. This helps us to avoid edges with very small variation in one dimension and a large variation in the other as opposed to $det(C_e)$. Finally, we want to find a tree $T = (U', E')$ which minimizes $\sum_{e\in E'} w_e$. This is analogous to the problem of Minimum Spanning Tree(MST) in graph theory which is solvable $O(2n^2 \log(n))$ using Kruskal algorithm. The graph structure is constructed using the above approach for each mixture component coming from pose clustering independently. Refer to Figure 2 for visualization of MST results.

4 Experimental evaluation

This section describes our experimental setup and presents a comparative performance evaluation of the proposed method. We report results for the detection of six animal classes in PASCAL VOC 2007 and 2010 datasets [5]. The considered classes (*bird, cat, cow, dog, horse, sheep*) are highly non-rigid and remain challenging for the current detection methods.

For each training image of an object we have annotated up to nine object parts roughly covering the surface of an object[1]. Each visible object part has been annotated with a bounding box as illustrated in Figure 3. Our method is not restricted to the types of parts used in this paper and can be easily extended

[1] For each of the six animal classes we have annotated the following parts. *bird*: beak, closed wings, head, legs, open wings, tail; *dog*: front legs, tail, head, frontal face, profile face, ears, back; *cat*: front legs, tail, head, frontal face, profile face, ears, back; *cow*: back legs, front legs, head, frontal face, profile face, horns, ears, torso; *horse*: back legs, ears, front legs, frontal face, head, profile face, tail, torso; *sheep*: back legs, ears, front legs, frontal face, head, horns, profile face, tail, torso.

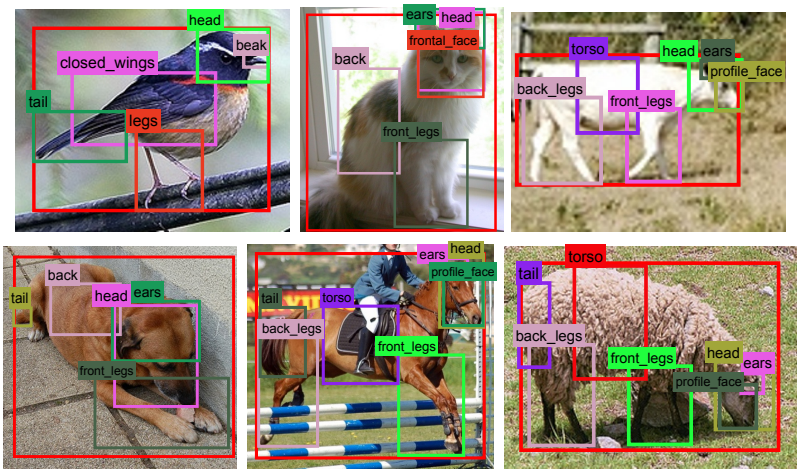

Fig. 3: Part annotation for training image examples of six animal classes.

to other parts and object classes. Part annotation has taken around eight person hours per one class in VOC 2007 dataset. Each object has been annotated by one person only[2].

Comparison with other methods. We compare results of our approach to the LSVM detector [1] and Poselets detector [22]. On the VOC 2007 dataset we follow [1] and apply post processing in terms of bounding box prediction to make our results comparable with results reported on the author's web-page [28] (see Table 1). For the VOC 2010 dataset (see Table 2) we retrain the detector [1] and obtain Poselets [22] detection results from the VOC challenge web-page. Overall our method improves LSVM baseline by **6.5%** and **5.1%** on VOC 2007 and VOC 2010 datasets respectively measured in terms of mean Average Precision (mAP) for six animal classes. We also obtain increase of performance by **2.7%** mAP compared to the Poselets method [22] which, similar to ours, uses additional part-level object annotation.

The APM method [24] uses part-level supervision and reports object detection results for *cat* and *dog* classes. Their evaluation setup on PASCAL VOC 2007 data, however, is not standard making comparison with [24] difficult. We note, however, that APM does not improve LSVM results for the *cat* while our improvement over LSVM is significant (the *cat* AP increases by 7%, see Table 1).

Detailed evaluation of proposed extensions. To better understand contributions of the proposed extensions, we report results for the four variants of our method and the LSVM baseline [1]. Results in Table 3 indicate a consistent improvement provided by the contributions of our approach. Figure 4 illustrates qualitative results of animal detection on a few example images from VOC 2007

[2] Annotations of parts are publicly available and can be found at `http://www.csc.kth.se/cvap/DPM/part_sup.html`

	bird	cat	cow	dog	horse	sheep	mAP
LSVM w/o context [1]	10.0	19.3	25.2	11.1	56.8	17.8	23.4
Ours	**12.7**	**26.3**	**34.6**	**19.1**	**62.9**	**23.6**	**29.9**

Table 1: Per-class results (Average Precision) for animal detection in VOC 2007.

	bird	cat	cow	dog	horse	sheep	mAP
LSVM [1]	9.2	22.8	21.2	10.4	40.8	27.0	21.9
Poselets [22]	8.5	22.2	20.6	18.5	**48.2**	**28.0**	24.3
Ours	**11.3**	**27.2**	**25.8**	**23.7**	46.1	**28.0**	**27.0**

Table 2: Per-class results (Average Precision) for animal detection in VOC 2010.

		bird	cat	cow	dog	horse	sheep	mAP
Supervised	LSVM [1]	10.1	18.4	24.7	11.3	57.8	18.6	23.5
	LSVM+Clus	9.1	26.1	29.8	12.8	56.0	24.5	26.4
	SParts	9.9	**26.5**	30.7	19.3	60.4	23.5	28.3
	Sparts+MST	11.8	23.5	32.6	**20.7**	61.0	**24.9**	29.1
	Sparts+MST+Occ	**12.5**	24.1	**34.4**	19.5	**62.2**	24.5	**29.5**
Semi Sup	Sparts+MST+Occ 75%	12.8	21.6	32.7	16.9	60.2	25.6	28.3
	Sparts+MST+Occ 50%	12.1	22.5	31.1	14.3	60.0	25.7	27.6
	Sparts+MST+Occ 25%	10.9	20.8	27.1	12.2	57.9	21.4	25.1

Table 3: Step-by-Step results for PASCAL VOC07 w/o bounding box prediction. **LSVM:** Release 4.0 implementation of [1] avaiable at authors webpage [28] **LSVM+Clus:** LSVM initialized with our pose clustering (see Section 3.2); **Sparts:** Supervised parts (instead of automatic LSVM parts) trained on top of pose clustering with a star model connectivity; **Sparts+MST:** Sparts with general acyclic connectivities using Minimum Spanning Trees; **Sparts+MST+Occ:** Sparts+MST with added occlusion filters and bias. Last rows show the results with partially annotated training data.

test set. Although we do not require part localization, correct part placement indicates the expected interpretation of the image by our method.

Semi-supervised learning. To reduce efforts for object part annotation, we investigate a semi-supervised learning setup where the part-level annotation is provided for a fraction of positive training samples only. As our method is able to handle samples with missing part annotation, we train detectors for $s\%$ positive training samples with part annotation and $(100-s)\%$ samples without part annotation. We compare detection results to the method trained on fully annotated samples. Results at the bottom of Table 3 and the corresponding plots in Figure 5 show the benefit of this semi-supervised learning setup provided by the use of samples with no part annotation. Moreover, even a fraction of part-annotated samples can result in significant improvements by our method.

Number of mixture components. We investigate the sensitivity of LSVM, LSVM+Clus and SParts methods (see definition in Table 5 captions) to the number of mixture components. Evaluation of mAP for these three methods

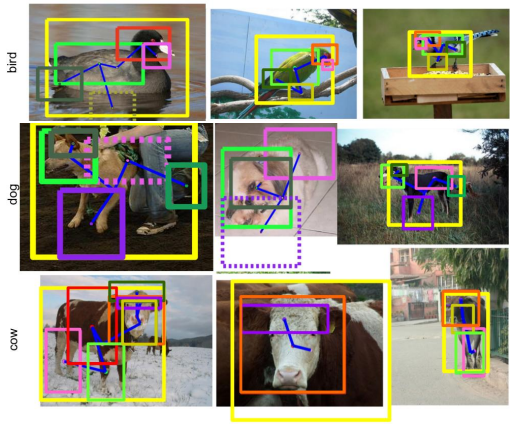

Fig. 4: Detection results for three animal classes. The detections of the part bounding boxes are overlaid with the tree structure of the model connecting the parts. Colors of part bounding boxes are consistent for different samples of the same class. Dashed bounding boxes indicates the activation of occlusion filters.

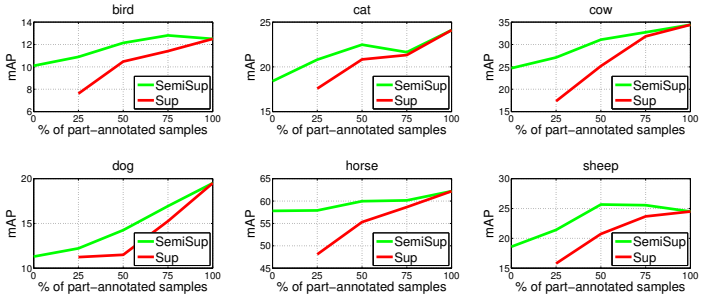

Fig. 5: Results of our model trained using different fractions of fully annotated samples. Red curves correspond to the training on s% samples with full part annotation. Green curves correspond to a semi-supervised training setup using X% fully-annotated samples as well as (100-s)% samples without part annotation. For each of the experiments (points in plot) 3 different random subsets are drawn. Mean average precision is computed over each three subsets.

and different numbers of components is illustrated in Figure 6. As can be seen, the proposed methods benefit from the increasing number of components while the performance of LSVM degrades. This demonstrates the room for modelling more intra-class variation when using more elaborated clustering criteria than the aspect ratio.

Latent positions. Although object parts are provided in our training by the annotation, we treat part locations as a constrained latent variables, i.e., we allow the location of parts to be optimized in the neighborhood of their annotated position. This is expected to reduce the influence of the imprecise part annotation and the possibly low discriminative power of manually annotated

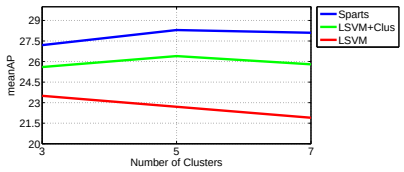

Fig. 6: Detection results for different number of mixture components obtained by LSVM, LSVM+Clus, and Sparts methods. Mean AP results are shown for 6 animal classes of PASCAL VOC 2007.

	bird	cat	cow	dog	horse	sheep	mAP
SParts	9.9	26.5	30.7	19.3	60.4	23.5	28.3
Sparts w/o Latency	1.1	10.2	16.2	0.8	44.4	8.5	13.5

Table 4: Detection performance on PASCAL VOC 2007 animals using either laten or fixed part positions.

object parts. To investigate the effect of latent positions, we conduct an experiment where we fix the positions of object parts to their annotated values. The comparison of detection results by our methods with and without latent part positions is summarized in Table 4 indicating the importance of the constrained latent variables in our method.

Part Localization. Our method localizes object parts as a bi-product of object detection. Here we evaluate object part localization on the VOC 2007 test set. We use the following three measures: (a)*Occlusion Detection (OD)* corresponding to the percentage of correctly recognized occlusion state[3] of the part, (b)*Percentage of Correctly localized Parts (PCP* [4]*)*[29] for non-occluded parts as well as (c)*Occlusion sensitive Percentage of Correctly detected Parts (OPCP)* for all parts defined as $\frac{OD*N_o+PCP*N_v}{N_o+N_v}$, where N_o and N_v are the number of samples with the part in the occluded or visible states respectively. It can be seen that (execpt for the part "head") our method is able to detect the occlusion state fairly well. We have observed low numbers of samples with heads being occluded. This should explain low OD values for the head for some objects.

Implementation details. We use HOG features and an optimization procedure adopted from the on-line implementation of [1] (vocrelease 4.0 [28]). Training our detectors takes about 20 hours on a recent 8-core Linux machine for one class on average. Detection takes about 4 seconds for a 300x500 image.
We cross-validated a few choices for the parameter W (pose clustering weight vector) and t_{ovp} (minimum part overlap constraint) for the class "cat" on VOC 2007 and fixed them for all classes with the same value. We set W to uniform

[3] We report a part occluded when either of the followings happen: a) the corresponding occlusion filter is activated b) a component is activated which does not include that part c) the part is detected with more than 70% of its area outside image

[4] We follow the approach of [10] when computing PCP in that we take all the detections in an image and for each ground truth object annotation we consider the detection with the highest score that has more than 50% overlap with annotation.

	head	front legs	fore legs	torso/back	tail
bird	21.4/25.4/25.3	-	30.6/12.1/18.2	-	55.4/6.1/19.9
cat	33.3/60.0/59.7	70.5/8.9/34.9	-	43.9/17.2/29.1	84.2/1.7/53.4
cow	27.3/36.3/35.8	65.0/25.9/34.6	57.5/37.1/42.5	84.2/58.2/66.5	-
dog	0.0/40.5/40.2	64.4/23.1/37.1	-	36.0/6.7/20.1	86.6/0.9/61.9
horse	66.7/65.7/65.7	30.3/37.3/35.8	65.2/39.3/46.7	91.3/57.7/66.0	48.3/32.0/39.6
sheep	26.1/29.4/28.9	65.5/17.6/34.3	78.6/10.9/33.7	76.5/57.1/65.1	80.8/2.4/61.4

Table 5: Part Detection performance evaluated on PASCAL VOC 2007, reported numbers are Occlusion Detection/PCP/OPCP respectively

weights and t_{ovp} to 30% overlap. We set LSVM penalty cost C to 0.002 as in original LSVM implementation [28]. We train a filter only for the parts that are visible in at least 30% of each cluster positive samples. Five mixture components are used for the final evaluations.

5 Conclusion

In this work we have explored the use of part-level supervision for the training of deformable part-based models. In particular, we have shown how to use part annotations (i) to derive better DPM initialization, (ii) to re-define the tree structure of the model and (iii) to model occlusions. Evaluation of these contributions has shown consistent improvement for the tasks of detecting animal classes in PASCAL VOC datasets. Our improvements in the semi-supervised setting indicate that even a fraction of part-level annotated samples can significantly improve the overall quality of the detector.

Acknowledgement. This work was supported by the VINST project funded by Swedish Foundation for Strategic Research (SSF), by the EIT ICT labs, by the Quaero program funded by OSEO and ERC grant VideoWorld. Computing resources were provided by the Swedish National Infrastructure for Computing (SNIC 001-12-16) via PDC.

References

1. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010) 1627–1645
2. Zhu, L., Chen, Y., Yuille, A., Freeman, W.: Latent hierarchical structural learning for object detection. In: CVPR. (2010) 1062–1069
3. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV. (2009)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. (2009)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: (The PASCAL Visual Object Classes Challenge 2010 VOC2010 Results)
6. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. TC **22** (1973) 67–92

7. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV **61** (2005) 55–79
8. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. (2006)
9. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: Retrieving people using their pose. In: CVPR. (2009)
10. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011) 1385–1392
11. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR. (2010) 2030–2037
12. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: BMVC. (2006)
13. Naderi Parizi, S., Oberlin, J., Felzenszwalb, P.: Reconfigurable models for scene recognition. In: CVPR. (2012)
14. Ott, P., Everingham, M.: Shared parts for deformable part-based models. In: CVPR. (2011)
15. Wang, Y., Tran, D., Liao, Z.: Learning hierarchical poselets for human parsing. In: CVPR. (2011) 1705–1712
16. Branson, S., Belongie, S., Perona, P.: Strong supervision from weak annotation: Interactive training of deformable part models. In: ICCV. (2011)
17. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) I:886–893
18. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. (2003) 1470–1477
19. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV **73** (2007) 213–238
20. Chen, Y., Zhu, L., Yuille, A.: Active mask hierarchies for object detection. In: ECCV. (2010)
21. Parkhi, O., Vedaldi, A., Jawahar, C.V., Zisserman, A.: The truth about cats and dogs. In: ICCV. (2011)
22. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: ECCV. (2010)
23. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR. (2011)
24. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: ICCV. (2011)
25. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR. (2012)
26. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
27. Harris, C., Stephens, C.: A combined corner and edge detector. In: Alvey Vision Conference. (1998)
28. Girshick, A., Felzenszwalb, P., McAllester, D.: LSVM Release 4 Notes. (http://www.cs.brown.edu/people/pff/latent/)
29. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)