Detection and Description of Objects by Detection of Body Parts

Anonymous CVPR submission

Paper ID 1052

Abstract

Detecting object parts individually is a difficult task, because often there are no strong appearance cues to distinguish them from the surrounding background or from other parts of the object. Detecting the whole object is also difficult, particularly for flexible articulated objects with high variations in object pose and partial occlusion. We propose a new method to combine object part detection with the detection of the whole object. We show that this gives better results than using the most successful whole object detectors (those evaluated on the Pascal Challenge) and also gives a richer description of the object in terms of parts that can be used for other applications.

In this paper, we apply our method to the six types of animal found in the Pascal VOC dataset (i.e. bird, cat, cow, dog, horse, and sheep). We first decompose these objects into their body parts (e.g., head, torso, legs, etc). This makes use of a new dataset of fully annotated object parts for Pascal 2010 dataset, which labels the positions and shapes of the parts. We evaluate our method on the animal classes of PASCAL 2010 dataset, and it demonstrates up to 9.4 improvement in average precision over the baseline approach (winner of PASCAL 2011 challenge).

1. Introduction

Detecting individual object parts is not an easy task in computer vision because of their great variability. For example, it is very hard to detect the heads of the animals shown in the second row of Figure 1. But the task becomes easier if we can detect the whole animal and use it as context. On the other hand, detecting the whole object may be hard due to occlusion, deformations, and pose variations. Most of the objects we see in our everyday life are partially occluded, which causes problems for computer vision algorithms designed to detect whole objects.

The current best object detectors are unable to reliably detect the objects shown in the top row of Figure 1. But using body parts as context could help make this task practical by, for example, detecting the heads of the animals. Conversely detecting the animal heads can sometimes be very difficult even with state of the art detectors, see bottom row of Figure 1, but it is possible to first detect the whole object and use it as context to detect the body parts.

Figure 1. Top row: in some situations it is hard to detect the whole object, but easier to detect the object parts (e.g., the head). Bottom row: in other cases, detecting the whole object is easier than detecting some of the body parts (e.g., the head).

In this paper, we address the problem of detecting the six types of animals that appear in the Pascal Challenge dataset. We represent them in terms of their body parts. We show that even if the accuracy for detecting individual body parts is not high, they can nevertheless help give a gain in improvement for detecting objects. We note that there is an extensive literature about part-based object detectors, where supervised [20, 17, 1] or unsupervised (latent) [21, 7, 15] parts have been used.

Our strategy is to use detectors for body parts, and the whole object, to yield a set of hypotheses for the presence of body parts and objects in the images and their properties (e.g., size, position, and appearance). The next stage is to detect the object by combining these hypotheses using consistency relations between them (e.g., their relative sizes, positions, and appearance). This requires us to decide which body part hypotheses should be combined with a whole body hypothesis. We must also take into account that some body parts may not be present in the image, or that our detector has failed to find them. We formulate this problem using a conditional random field (CRF) defined over a

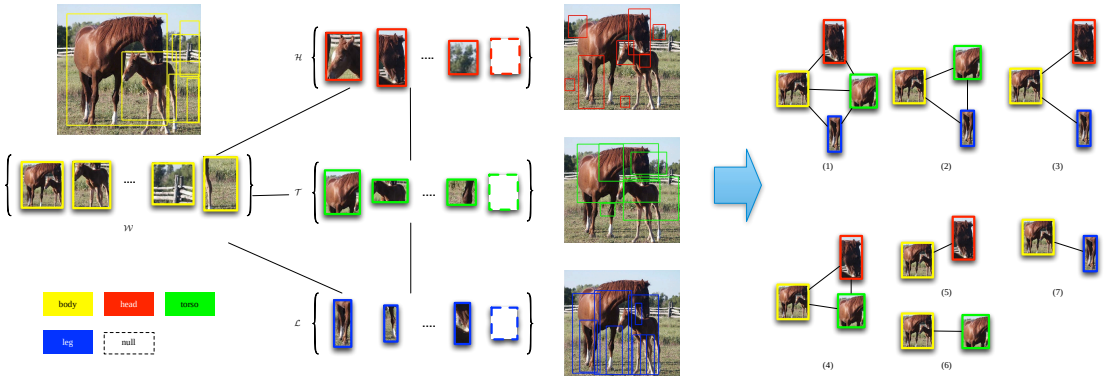

Figure 2. The graphical model for the horse. We have seven models corresponding to different combination of body parts and body.

graph with nodes representing the body parts and the whole body, see Figure 2. The state variable of a body part node indicates which, if any, of the hypotheses of that object are correct and specify their properties. The edges of the CRF specify consistency relations between different body parts (and the whole body) requiring, for example, similar scale and appearance (e.g., the head and torso of a cat are likely to have similar colors). We perform inference to estimate the most probable (i.e. lowest energy) state of the CRF and output the lowest energy as our confidence measure for detecting the object (in practice, we discovered a greedy algorithm which sped up computation).

Our method shows significant improvement (up to 9.4%) over the base code of the co-winner of PASCAL VOC 2011 detection challenge [4]. We note that our method also outputs the positions of body parts, while the base code only outputs a bounding box for the whole object. We used supervised learning for training the body part detectors. For this purpose, we annotated object parts for all of the images of animal categories of PASCAL VOC 2010 (see the experiments section for more details).

The paper is organized as follows. In Section 2, we describe our model for combining body parts and whole. We define a conditional random field by specifying an energy function that captures a set of consistencies between parts and body hypotheses. Also we propose a simple greedy strategy to find the minimum energy configuration of the model. In Section 3, we provide more details about the dataset labels, which includes annotations for the parts of PASCAL VOC categories. Also, we explain the details of our experiments and show we obtain a significant gain over the baseline object detector, which was one of the state of the art methods as evaluated in the PASCAL VOC 2011 challenge.

Related work: There is a considerable body of work on part-based object detectors which encode parts using latent variables [5, 21, 6, 7, 15, 11]. Typically these methods formulate object detection within a discriminative framework and so learn parts that are most effective for discriminating the object from the background clutter of the images. Hence many of these methods are not suitable for recovering the body parts of the object.

Most work on human pose estimation in recent years has relied on anatomical body parts [16, 14, 18, 20]. These methods can be divided into two major categories: approaches that learn the models for parts and body simultaneously and the ones that learn the models for parts independently from the body model. Our approach is close to the latter category as we learn body and part models separately and then combine them together to boost object detection performance. One disadvantage of simultaneous learning of part and body models is that the learning becomes more difficult because of more local minima in the model.

There are fewer works dealing with the body parts of other object classes [1, 17, 2, 3]. Strong supervision is used in these approaches for configuration clustering and occlusion reasoning. On average, our results are significantly better than the results of these approaches. We cannot directly compare our results with some of them as we work on the full PASCAL 2010 dataset not a selected subset.

2. The Model

This section describes our model. Subsection (2.1) describes how we generate hypotheses for body parts and for the whole object in each image.

2.1. Obtaining the Body Part and Whole Body Hypotheses

The input to our model for each image is a set of hypotheses for each body part and the whole object, see Figure 3. These were obtained by using a hierarchical latent-SVM classifiers which have been successfully applied in the Pascal Object Detection Challenge [4]. These classifiers were applied to the image at a range of scales and positions. There are only two differences to the standard procedure for Pascal. Firstly, classifiers are trained separately

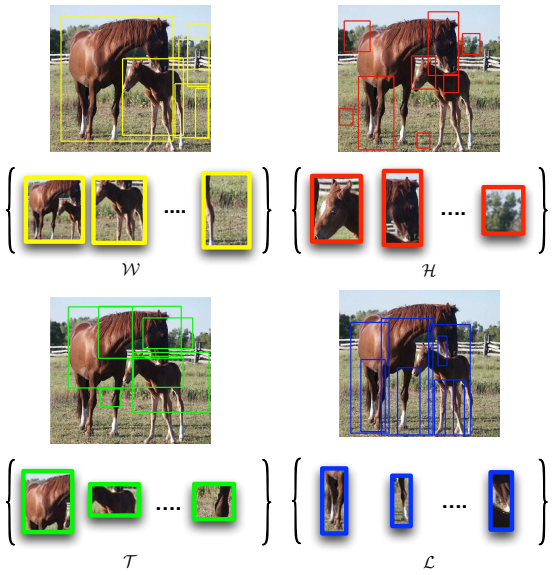

Figure 3. The set of hypotheses for body parts and whole body ($\mathcal{P}$, $\mathcal{T}$, $\mathcal{L}$ and $\mathcal{W}$) are shown.

for each body part and also for the whole object. Secondly, we threshold the classifier scores for each body part and the whole object so as to ensure that we have a limited number of detections in the image (typically these thresholds resulted in less than fifteen responses in each image). The thresholds were chosen to ensure that there were few false negatives and a limited number of false positives in each image.

More formally, for an image $\mathbf{I}$, let $\mathcal{W} = \{w_1, \ldots, w_k\}$ be the set of detection hypotheses obtained using whole-body detector. Let $\mathcal{H} = \{h_0, h_1, \ldots, h_n\}$, $\mathcal{T} = \{t_0, t_1, \ldots, t_m\}$, and $\mathcal{L} = \{l_0, l_1, \ldots, l_p\}$ be the head, torso, and leg hypotheses. Here h_0, t_0, l_0 are dummy variables representing (respectively) the null hypotheses that the head, the torso, or the leg are not present, or detected, in the image. Any other (i.e. real) hypothesis d has a *score* $s(d)$ which is given by its classifier. Each hypothesis also has a *geometric attribute*, which is the bounding box $B(d)$ (also output by the classifier). From this we can compute other geometric properties such as the size $|B(d)|$ of the hypothesis or the position of its center. Each hypothesis also has an *appearance attribute* $A(d)$, for example the color histogram within $B(d)$. We will sometimes use $\mathcal{P}$ as shorthand to refer to all the body part hypotheses (e.g., $\mathcal{H}, \mathcal{T}, \mathcal{L}$.) We will sometimes use $\mathcal{P}_\tau$ to refer to the hypotheses for a body part τ, where τ stands for head, torso, or leg.

2.2. Graphical Model of each Object

We now define a graphical model for each object. This has four nodes representing (respectively) the whole body, the head, the torso, and the legs. The states of the node represent the hypotheses for the corresponding body part (or whole body) and include a null hypothesis which allows for the body part to be undetected. The states are specified by $\mathbf{x} = (w, h, t, l)$ which take values within the hypothesis sets $\mathcal{W}, \mathcal{H}, \mathcal{T}, \mathcal{L}$ respectively (note our definition allows the state variables to represent the null hypotheses). Each hypothesis d (except the null hypotheses) has a geometric attribute B, and an appearance attribute A (as described in the previous subsection).

Next we define a probability distribution – a conditional random field – for the configuration $\mathbf{x}$ of the graphical model. The graphical model for the horse is illustrated in Figure 2. There are eight different graphical structures to allow for the possibility that one or more body parts are missing. Each graph structure has binary terms connecting all of the nodes except the leg and the head. We show seven of the graphical models in Figure 2 (where we ignore the case when none of the body parts is detected).

We express the distribution as a Gibbs distribution which requires us only to specify an energy function. This energy is composed in terms of potentials for relations between body parts and the whole body. This takes the form for the model which includes all body parts (and the other six models are similar):

$$P(\mathbf{x}|\mathcal{W}, \mathcal{P}) = \frac{1}{Z} \exp\{-E(\mathbf{x}|\mathcal{W}, \mathcal{P})\}. \quad (1)$$

The model is conditioned on the set of whole object hypotheses $\mathcal{W}$ and body part hypotheses $\mathcal{P} = \mathcal{H}, \mathcal{T}, \mathcal{L}$ detected in the image.

Observe that many hypotheses are generated in each image for each body part and for the whole object. The state of each graph node can take any one of these hypotheses. For instance, if the head and torso detectors generate 10 head hypotheses and 15 torso hypotheses for an image, then the head and torso nodes can take 10 and 15 possible states respectively.

For each image, and object model, we find the best associations between the hypotheses for the whole object and the body parts by solving the following optimization problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x}|\mathcal{W}, \mathcal{P}), \quad (2)$$

,where $\mathcal{W}$ and $\mathcal{P}$ are the set of body and part hypotheses for the image.

The next subsections define the potential terms for the model and describes the inference algorithm used to compute $\mathbf{x}^*$. The potentials include terms which relate the body parts to the whole object, and the head to the torso and the torso to the legs.

2.3. Potential terms relating the body parts to the whole object

We now describe the potential terms which constrain the relationship between body parts and the whole object. A whole body hypothesis may overlap with several head hypothesis for example, but at most one of them can be correct. These potential terms encode the probability of assigning a body part to the whole object based on geometric consistency. Recall that the set of whole object hypotheses is $\mathcal{W} = \{w_1, w_2, \ldots, w_k\}$. The set of real hypotheses (i.e. not including the null hypothesis) for body part τ is given by $\mathcal{P}_\tau = \{p_{\tau 1}, p_{\tau 2}, \ldots, p_{\tau P_\tau}\}$, where the number of hypotheses for τ is given by P_τ.

In the following paragraphs, we use $|.|$ to denote the size of a bounding box and $\cap$ to denote the intersection of the bounding boxes. More specifically, $|d_i| = |B(d_i)|$ is the size of the bounding box of d_i, $|d_i \cap d_j| = |B(d_i) \cap B(d_j)|$ is the size of the overlap between the bounding boxes of d_i and d_j.

Overlap potentials: Ideally, the bounding box for a part should be completely inside the whole body bounding box. However, since our detectors are not perfect, we cannot expect a full overlap between bounding boxes. There is a high probability that part detectors generate boxes that are inside the body bounding box. Our goal is to encourage such boxes. On the other hand, if a part box does not overlap with the body box, there is a high chance that the part does not belong to that body. This motivates us to define the potential $\psi_{ov}(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P})$ relating the state $p_{\tau i}$ of a body part (i.e. head, torso, or leg) to the state w_j of the whole object by:

$$\psi_{ov}(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P}) = \min(C_O, \frac{|p_{\tau i} \cap w_j|}{|p_{\tau i}|})/C_O, \quad (3)$$

where C_O is a truncation parameter allowing some robustness against the localization noise in the part and body detectors. We set $C_O = 0.9$ for all of our experiments (recall that $|p_{\tau i} \cap w_j|$ denotes the intersection of the bounding boxes of these two hypotheses).

Scale potential: There is often a relationship between the scale of the body parts and the whole object. If one body part is very close to the camera and whole body is farther away then the scale relationship is violated, but in general this is a rare case in the images we observe. We model the relative scale of parts and body with a Gaussian distribution and use it as another term in our model. The following equation specifies the scale potential, ψ_S for body part hypothesis $p_{\tau i}$ and whole object hypothesis w_j:

$$\psi_S(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P}) = \max(0, C_S - \frac{\left\| \frac{|p_{\tau i}|}{|w_j|} - \mu_S \right\|}{\sigma_S})/C_S. \quad (4)$$

The mean and variance of the scale ratio distribution are denoted by μ_S and σ_S, respectively and learned from training data by simple maximum likelihood estimation. These parameters are learned for each object class separately. We also have a truncation parameter, C_S, for this term, which means that the penalty for the values that are very far from the mean of the distribution becomes constant after a certain point. We use $C_S = 4$ for all of the experiments in this paper. Here, $\|.\|$ is the $L2$ norm.

Finally, we weight the overlap and scale potentials by the scores $s(w_j), s(p_{\tau i})$ of their respective whole object and body part hypotheses. Hence $\psi_{ov}(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P}) \mapsto s(p_{\tau i}) s(w_j) \psi_{ov}(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P})$ and $\psi_S(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P}) \mapsto s(p_{\tau i}) s(w_j) \psi_S(p_{\tau i}, w_j | \mathcal{W}, \mathcal{P})$. The scores are not calibrated but calibration methods such as [19] might produce even better results. The intuition for this weighting is that if we are very confident about an object hypothesis, its associated part hypotheses should receive more confidence and vice versa i.e. body and part hypotheses provide context for each other. Figure 4 show different terms we have used to compute the potentials.

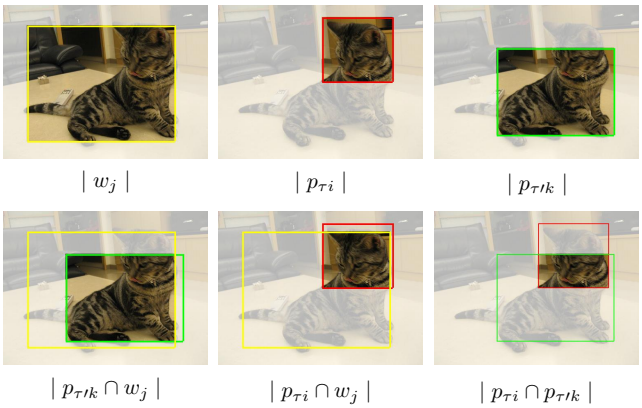

Figure 4. The geometric quantities used for computing the relations between body part and whole object hypotheses. These include the size of the bounding boxes of the hypotheses, which are used to compute scale relations, and the intersections of the bounding boxes (used to impose constraints, that the head of the cat is likely to mostly be inside the bounding box of the whole object).

2.4. Potentials relating the Body Parts

We use pairwise terms to enforce consistency between the body parts. Note that these terms are only imposed between the head and the torso, and between the torso and the legs (there is no direct connection between the head and the legs). These are similar to those described in the previous subsection, except that they also include appearance cues.

Pairwise scale potential: There is a relationship between the scales of different body parts. We define the scale consistency for body part hypotheses $p_{\tau i}$ and $p_{\tau' j}$ similar to

the scale term for body parts and the whole object in the previous subsection (but this term only applies to hypotheses from different body parts – e.g., between head and torso hypotheses or torso and leg hypotheses).

$$\phi_S(p_{\tau i}, p_{\tau' j}|\mathcal{P}) = \max(0, C_S - \frac{\left\| \frac{|p_{\tau i}|}{|p_{\tau' j}|} - \mu_{\tau\tau'} \right\|}{\sigma_{\tau\tau'}})/C_S, \quad (5)$$

where, as before, p's are the body part hypotheses. C_S is a truncation parameter (similar to that in the previous subsection). $\mu_{\tau\tau'}$ and $\sigma_{\tau\tau'}$ are the parameters of the Gaussian distribution, which are learned from the training data.

Appearance similarity potential: The body parts that belong to the same object tend to have similar appearance. At least, this is a reasonable assumption for animals. It is unlikely that the color or texture of the head of an animal is totally different from the color or texture of its torso (although this assumption is less true for birds). In the current implementation we use color cues only.

The appearance attribute for body part hypothesis $p_{\tau i}$ is specified by $A(p_{\tau_i})$. In this paper, it is specified by the color histogram $\mathcal{H}_{\tau i}$ within the bounding box $B(p_{\tau i})$. The color histogram that we use in the experiments has 48 bins, where each color channel corresponds to 16 bins. We compute the histograms in RGB space. The rationale for the appearance similarity potential is that two body parts that have the same color are more likely to belong to the same object. The appearance potential between body parts hypotheses $p_{\tau i}$ and $p_{\tau' j}$ is defined as follows:

$$\phi_C(p_{\tau i}, p_{\tau' j}|\mathcal{P}) = -\mathcal{X}^2(\mathcal{H}_{\tau i}, \mathcal{H}_{\tau' j}), \quad (6)$$

where $\mathcal{X}^2(.,.)$ is the chi-squared distance between the color histograms.

We have explored some alternatives for the color term, but the one discussed above performed best for us. For instance, we tried obtaining segments inside each body part hypothesis by the method of [9] and used color similarity between the most similar segments of two body parts. But the overall performance was lower than if we simply used the color histogram over the whole bounding box of the body part.

Body Part overlap: Another cue that we use in our model is the amount of overlap between body parts. If two body part hypotheses are totally separated, then there is a chance that they belong to different objects, while overlapping body part bounding boxes tend to belong to the same object with higher probability. We denote pairwise part overlap as ϕ_{ov} and define it as:

$$\phi_{ov}(p_{\tau i}, p_{\tau' j}|\mathcal{P}) = \min(C_O, \frac{|p_{\tau i} \cap p_{\tau' j}|}{|p_{\tau i}|})/C_O, \quad (7)$$

where C_O is the same truncation parameter that we defined before for part-body overlap term.

In the next subsection, we define the energy function built from these terms. It should be noted that we do not use spatial relationship between body parts (body parts can be at any orientation with respect to each other so it is not a strong cue for our model). We note that our hypotheses for body parts and whole objects are obtained using hierarchical latent-SVM models which encode spatial relations between latent parts (but the "parts" used in hierarchical latent-SVM models are hard to interpret for animals and certainly do not correspond to body parts).

2.5. The Full Model

We now put all the potential terms together to get the energy function for the model. This is defined by:

$$E(\mathbf{x}|\mathcal{W}, \mathcal{P}) = -\left\{ \sum_{\tau,i,j} \psi(p_{\tau i}, w_j|\mathcal{W}, \mathcal{P}) + \sum_{\tau,\tau',i,j} \phi(p_{\tau i}, p_{\tau' j}|\mathcal{P}) \right\}, \quad (8)$$

where,

$$\begin{aligned} \psi(p_{\tau i}, w_j|\mathcal{W}, \mathcal{P}) &= \lambda_1 \psi_{ov}(p_{\tau i}, w_j|\mathcal{W}, \mathcal{P}) + \lambda_2 \psi_S(p_{\tau i}, w_j|\mathcal{W}, \mathcal{P}) \\ \phi(p_{\tau i}, p_{\tau' j}|\mathcal{P}) &= \lambda'_1 \phi_S(p_{\tau i}, p_{\tau' j}|\mathcal{P}) + \\ &\quad \lambda'_2 \phi_C(p_{\tau i}, p_{\tau' j}|\mathcal{P}) + \lambda'_3 \phi_{ov}(p_{\tau i}, p_{\tau' j}|\mathcal{P}). \end{aligned} \quad (9)$$

The parameters λ are the weights for each term in the energy function. We provide more details about them in the experiments section. We set these parameters based on a set of validation images for the category of interest, but they could also be learned with the existing parameter learning methods for CRFs (for example, [10]).

We also allow for the possibility that some, or all, body parts are missing (i.e., if the state variable takes the null hypothesis for that body part). This is done by paying a penalty in the energy for each body part that is missing. In addition, we drop all potential terms in the energy which involve that body part (this is equivalent to performing inference over different graphical models simultaneously).

2.6. Inference

To detect the optimal configuration for each object, we need to find the optimum state of Equation 8 by solving $\mathbf{x}^* = \arg\min E(\mathbf{x}|\mathcal{W}, \mathcal{P})$. This finds the best associations between the body parts and the whole object hypotheses. It outputs a configuration $\mathbf{x}^*$ of the object. Its energy $E(\mathbf{x}^*|\mathcal{W}, \mathcal{P})$ is used as a measure of confidence for the detection of the object.

We studied several algorithms for solving for $\mathbf{x}^*$. Exhaustive search is possible since, although the number of possible state configurations grows exponentially with the number of body parts, the number of body parts is small in our experiments (e.g., head, torso, legs) and the number

of hypotheses for each body part is limited (typically less than 15. This would only require evaluating at most a few $1,000$'s of configurations for each image, which is certainly feasible. But we preferred a faster algorithm because we want to test and evaluate our approach on very large numbers of images, of the order of $10,000$'s. We can improve over exhaustive search by exploiting the graph structure of the model and indeed we implemented TRW-S [12], but this still took time when evaluated on large datasets.

Instead we discover a greedy heuristic which performed as well for this problem as TRW-S but was an order of magnitude faster. This was based on the observation that each whole object hypothesis was usually strongly associated with a very limited number of body parts. We could therefore prune out most of the possible configurations by thresholding $\psi(.|\mathcal{W}, \mathcal{P})$ to obtain a limited set of whole object and body part pairs. Then, for each whole body hypothesis we can simply enumerate the limited number of configurations and compute the one with lowest energy. Then we take the minimum over all whole body hypotheses.

3. Experiments

Dataset: We use the PASCAL 2010 dataset for training and testing. We focus on the six animal categories because those are hard to detect due to their highly flexible body structure. For training, we have supplemented PASCAL by providing labeling for the masks/bounding boxes of object parts. We note that this is a different annotation than that used in [1, 17].[1] Figure 5 shows the annotations for some example categories.

In this section, we provide the details for our experiments and show that our method provides significant boost over the baseline method, which is one of the current state-of-the art methods.

Our baseline is the base code of the winner of PASCAL 2011 detection challenge [4] [2]. By base code, we mean it does *not* include post-processing stages such as context re-scoring, bounding box prediction, etc. (which typically add 3-4 AP improvement). It should be noted that any other object detector can be used instead, and our method does not assume a specific choice of baseline.

The baseline detector provides a set of hypotheses for body parts and whole object. We train our model for each body part separately using the corresponding bounding boxes and then run inference on the test images to obtain the required hypotheses for each body part. These body part hypotheses are the possible states of the nodes of our graphical model. The base detector uses unsupervised (latent) parts, but there is no relation between these latent parts and our body parts.

[1]This labeling is not yet publicly available, but it is being submitted for publication and, in any case, will be made available if this paper is accepted.

[2]The code is not publicly available. We obtained it through personal communication.

	Bird	Cat	Cow	Dog	Horse	Sheep	mAP
Baseline[4]	14.4	40.0	22.3	27.7	36.3	29.1	28.3
Ours	**16.0**	44.3	**26.3**	31.7	45.7	**31.0**	**32.5**
gain	1.6	4.3	4.0	4.0	9.4	1.9	4.2
Poselets [2]	8.5	22.2	20.6	18.5	**48.2**	28.0	24.3
Sup-DPM [1]	11.3	27.2	25.8	23.7	46.1	28.0	27.0
DisPM [13]	-	**45.3**	-	**36.8**	-	-	-

Table 1. Average Precision for full body detection. Note that, unlike us, [13] uses context cues (which typically yield improvements on 3-4 AP).

	Bird	Cat	Cow	Dog	Horse	Sheep	mAP
w/o body-part scale	15.3	41.9	23.4	30.8	**47.9**	30.9	31.7
w/o body-part overlap	13.9	41.6	19.3	26.6	35.8	20.4	26.3
w/o part-part	**16.1**	**44.3**	26.0	31.5	45.4	30.7	32.3
w/o change box	15.7	42.5	25.6	**32.2**	45.1	29.8	31.8
all	16.0	**44.3**	**26.3**	31.7	45.7	**31.0**	**32.5**

Table 2. Effect of each term in the model

		Bird	Cat	Cow	Dog	Horse	Sheep	mAP
head	wo model	5.5	48.1	21.8	11.1	41.3	8.6	22.7
	in model	5.7	42.5	18.3	9.4	31.3	5.3	18.8
torso	wo model	3.0	10.4	12.9	1.9	29.2	12.5	11.7
	in model	7.1	14.7	16.8	2.2	31.1	14.8	14.5

Table 3. Average Precision for part detection

To generate whole object hypotheses, we performed training on `trainval` set of PASCAL VOC 2010 and tested on the `test` portion of the dataset. For training the body part models, we use the same split of our dataset of part annotations. For our experiments, we use $\tau = 1, 2$, for head and torso. Our method can also handle other body parts, as we show in Figure 6, but they do not increase the overall performance of object detection.

The final result of our method is a set of bounding boxes for the full-object and its constituent parts. We used two approaches for generating the final bounding boxes. In the first approach, we use the whole object bounding boxes that the baseline detector outputs. These are the original hypotheses that we use in the model. The only difference is that their scores will be different, because they are modified by the detection of the body parts. In the second approach, we choose the average of the original whole object bounding box and a box that we obtained by putting a box around the body parts associated to that particular whole object. As shown in Table 2 (4th row), the average AP for the second approach is slightly higher than that of the first approach.

To evaluate the performance of our full-object detector, we also need to assign a score to each body part bounding box. The score for each body part bounding box is the negative of the energy that we obtain from Equation 8.

We use the standard PASCAL 50% intersection over union overlap (IOU) criterion to evaluate the performance

Figure 5. Example annotations of the dataset. We use the bounding box around the mask for training and evaluation.

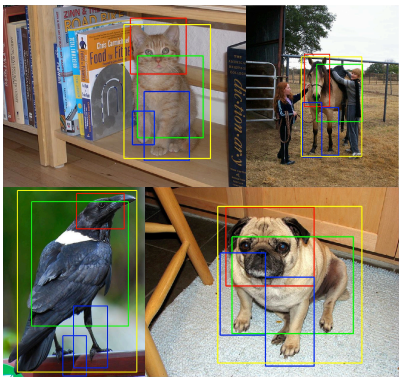

Figure 6. Output of our method: whole body (yellow), head (red), torso (green), legs (blue)

of our full-object detector. We use Average Precision as the evaluation criteria. As shown in Table 3, we provide significant improvement over the baseline method for all six of the animal categories. The highest gain is for horse category, where the improvement in average precision is 9.4%. In addition, unlike the baseline method, we also provide information about the body parts of the full-object.

We also show the results for Poselet [2], [1], and [13] which are also considered as strongly supervised methods. However, our method is not directly comparable with theirs as we have used a different set of images for training. Note that unlike [1, 2], we do not use any constraint about the relative spatial location of parts. Also, unlike [13] we do not use context rescoring.

We also evaluate body part detection performance. The output of our method includes a set of part bounding box hypotheses as well as their body association label. We use the same criteria as above (50% intersection over union) to measure the accuracy of body part detection. Other methods such as PCP [8] or criteria of [17] could be used for the evaluation, but we choose this criteria so we can provide a comparison with full-object detection accuracy.

Table 3 shows the average precision for body part detection without using the model and also after processing with the model. Note that the body part detection performance in isolation is often lower than the full-body detection performance. We expected that performance of part detection inside the model to be lower since we ignore the parts that are not associated to any body, but interestingly, the model improves torso detection performance for all categories.

We also show how each term in the model affects object detection performance. For this purpose, we set the weight for each term to zero one at a time and measure the accuracy. Table 2 shows the results where one term is missing from the model. For instance, for the horse category, ignoring the scale term improves the results. That might be due to the fact that for horses having a mixture of scale ratios is more suitable.

The weights λ in Equation 9 are set as follows for all of the classes: $\lambda_1 = 0.3$, $\lambda_2 = 0.7$, $\lambda'_1 = 0.4$, $\lambda'_2 = 0.3$, and $\lambda'_3 = 0.3$. Figure 7 shows the output of our method.

4. Conclusion

We described a method for improving the detection of objects by using body part detectors in conjunction with object detectors. We use a baseline detector to output hypotheses for body parts and the whole object which are combined using a conditional random field. We demonstrate that this method outperforms the baseline object detector by 4.2% on average when evaluated on the PASCAL VOC 2010 dataset. We also evaluate the ability of our approach to detect body parts and show that it improves their detectability.

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, 2012. 1, 2, 6, 7

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2, 6, 7

[3] S. Branson, P. Perona, and S. Belongie. Strong supervision from weak annotation: Interactive training of deformable part models. In *ICCV*, 2011. 2

[4] Y. Chen, L. Wan, L. Zhu, R. Fergus, and A. Yuille. In *The PASCAL Visual Object Classes Challenge Workshop*, 2011. 2, 6

[5] D. Crandall and D. Huttenlocher. Weakly supervised learning of part based spatial models for visual object recognition. In *ECCV*, 2006. 2

[6] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. In *ECCV*, 2008. 2

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. 32(9), 2010. 1, 2

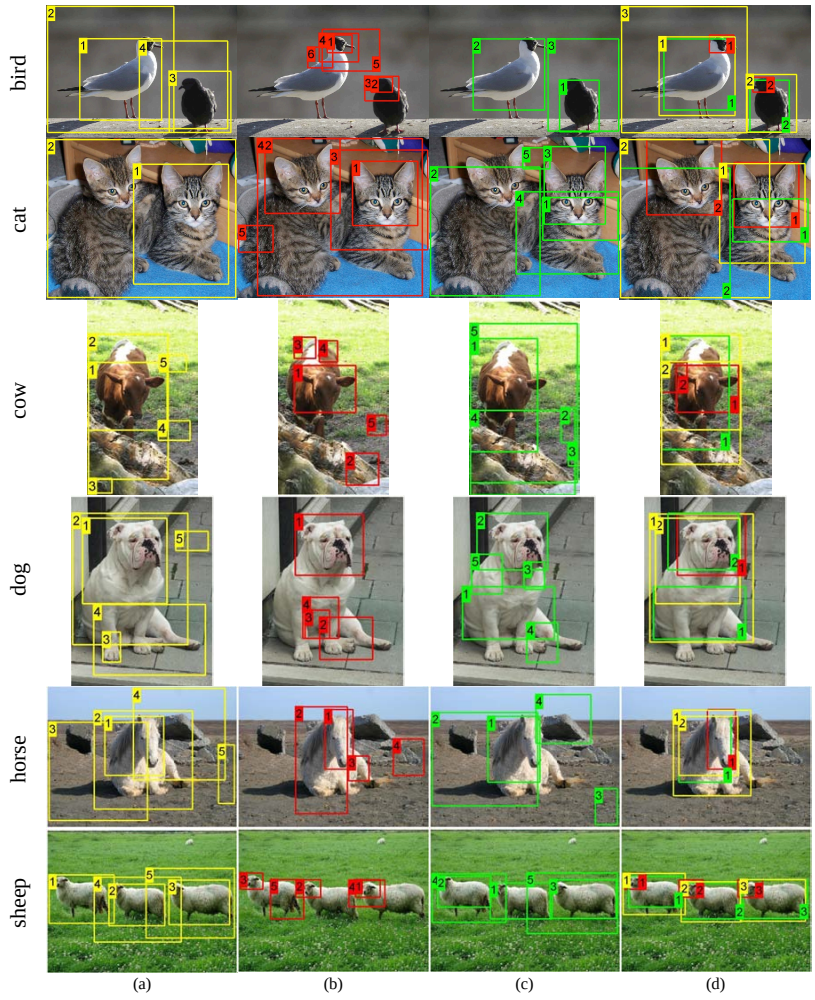

Figure 7. (a) Whole body hypotheses, (b) Head hypotheses, (c) Torso hypotheses, (d) Final output of our method. The number inside each box shows the rank for that box.

[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008. 7

[9] M. Galun, E. Sharon, R. Basri, and A. Brandt. Texture segmentation by multiscale aggregation of filter responses and shape elements. In *ICCV*, 2003. 5

[10] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 5

[11] L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *CVPR*, 2010. 2

[12] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006. 6

[13] O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The truth about cats and dogs. In *ICCV*, 2011. 6, 7

[14] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 2

[15] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent crfs. In *CVPR*, 2010. 1, 2

[16] V. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *ECCV*, 2010. 2

[17] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV*, 2011. 1, 2, 6, 7

[18] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *ECCV*, 2010. 2

[19] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object models for image segmentation. *PAMI*, 2011. 4

[20] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In *CVPR*, 2011. 1, 2

[21] L. Zhu, C. Lin, H. Huang, Y. Chen, and A. Yuille. Unsupervised structure learning: Hierarchical recursive composition, suspicious coincidence and competitive exclusion. In *ECCV*, 2008. 1, 2