FUSING GENERIC OBJECTNESS AND DEFORMABLE PART-BASED MODELS FOR WEAKLY SUPERVISED OBJECT DETECTION

Yuxing Tang[*1], *Xiaofang Wang*[*1], *Emmanuel Dellandrea*[1], *Simon Masnou*[2], *Liming Chen*[1]

[1]Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France
[2]Université Lyon 1, ICJ, UMR5208, F-69622, France

ABSTRACT

In the context of lack of object-level annotation, we propose a model that enhances the weakly supervised deformable part model (DPM) by emphasizing the importance of size and aspect ratio of the initial class-specific root filter. For each image, to extract a reliable bounding box as this root filter estimate, we explore the generic objectness measurement to obtain a reference window based on the most salient region, and select a small set of candidate windows by adaptive thresholding and greedy Non-Maximum Suppression (NMS). The initial root filter estimate is decided by optimizing the score of overlap between the reference box and candidate boxes, as well as their corresponding objectness score. Then the derived window is treated as a positive training window for DPM training. Finally, we design a flexible enlarging-and-shrinking post-processing procedure to modify the output of DPM, which can effectively fit to the aspect ratio of the object and further improve the final accuracy. Experimental results on the challenging PASCAL VOC 2007 database demonstrate that our proposed framework is effective and competitive with the state-of-the-arts.

Index Terms— Object detection, weakly supervised learning, deformable part-based models, objectness, post-processing

1. INTRODUCTION

Object detection/localization in images is one of the most widely studied problems in computer vision. For most of the existing methods, a fully supervised learning (FSL) approach is adopted [1, 2], where positive training images are manually annotated with bounding boxes encompassing the objects of interest. However, manual annotation for large-scale image database is extremely laborious and unreliable [3]. As a result, in contrast to the traditional FSL, there has been a great interest in weakly supervised learning (WSL) for object detection [4, 5, 6, 7, 8, 9, 10], where the exact object locations in positive training examples are not provided, given only the binary labels indicating the presence or absence of the objects.

Deformable Part-based Models (DPM) [2] and its variants [11, 12], are the leading technique to object detection with full supervision on the challenging PASCAL VOC datasets [13]. The DPM represent an object with a coarse root filter that approximately covers an entire object and several higher resolution part filters that cover smaller parts of the object. In the standard (fully supervised) DPM framework, the positive ground-truth object bounding boxes are treated as the initial root filters, and it is allowed to move around in its small neighborhood to maximize the filter score. The locations of parts are treated as latent information as the annotations for parts are not available. Megha *et al.* [5] modify the fully supervised DPM to a weakly supervised one, without object-level annotations, by treating the location of root filter and part filters full latent, and learning structural object detectors based on the entire image (root filter location is initialized randomly based on a window which has at least 40% overlap with the positive training image, and its aspect ratio is initialized roughly to the average of the aspect ratios of positive training examples). However, the specific size and location of the initial root filter, as well as their aspect ratio are indicated to have a significant impact on the final localization result [1, 2, 5]. And to our best knowledge, methods for initializing the root filter as well as the definition of the aspect ratio of the objects in weakly supervised DPM, have not been well studied in [5].

To take advantage of the outstanding object detection performance of fully supervised DPM, in this paper, we propose a model enhancing the weakly supervised DPM by emphasizing the importance of location and size of the initial class-specific root filter. To be precise, we explore the objectness approach [14], which generates class-independent object proposals with corresponding scores to their probabilities of being object windows, and adaptively extract a reliable window from the derived object proposals for each image as the initial root filter estimate for training DPM detector. Finally, a flexible enlarging-and-shrinking post-processing procedure is proposed to modify the predicted output of DPM detector, which can effectively generate more accurate bounding boxes by better conserving foreground and cropping out plain

* indicates equal contribution to this work. This work was partially supported by French Research Agency, Agence Nationale de Recherche (ANR), through the Visen project, under the grant ANR-12-CHRI-0002-04, within the framework of the ERA-Net CHIST-ERA.

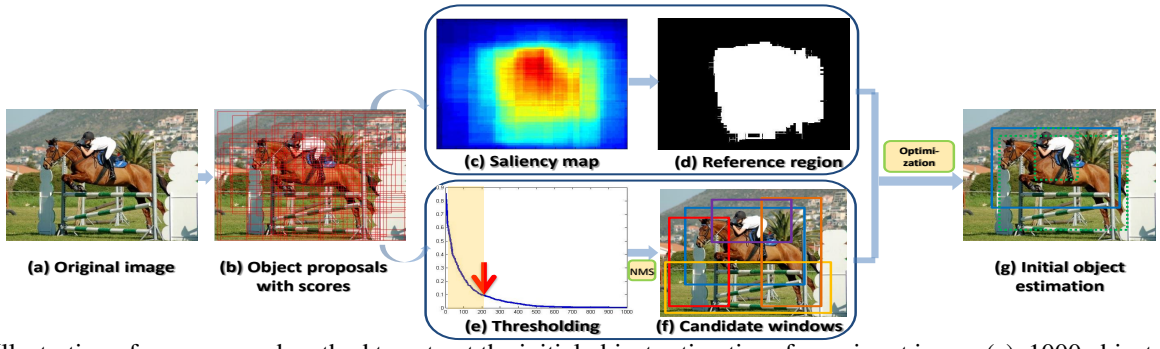

Fig. 1. Illustration of our proposed method to extract the initial object estimation: for an input image (a), 1000 object proposals (b) are sampled with corresponding scores to their probability to have object inside via the objectness measurement. (c) is the saliency map derived from (b), and (d) is the reference region obtained by thresholding (c). A finer set of candidate windows (f) are selected on the sorted proposals (e) by NMS. The blue window in (g) is our initial object estimation obtained by optimizing the overlap between (d) and (f).

background regions. Experimental results on the challenging PASCAL VOC 2007 database demonstrate that our proposed framework is effective for initialization of root filter, and shows competitive final localization performance with the other weakly supervised object detection methods[5, 10].

The rest of the paper is organized as follows: we present our method to extract reliable initial root filter for weakly supervised DPM and our technique to post-process the bounding box in Section 2, and in Section 3 we present our experimental results and the comparison with other methods on PASCAL VOC 2007 datasets. In Section 4, we conclude our work.

2. OUR APPROACH

In this section, we present our approach for improving the performance of DPM for weakly supervised object detection. In particular, we explore objectness measurement [14], which has been widely applied for various purposes in computer vision, to generates category-independent object proposals with corresponding scores to their likelihood of being object bounding boxes, and adaptively extract a faithful window from the derived object proposals for each image as the initial root filter size and position for DPM detector. We then briefly describe the training and detecting procedures with DPM. Finally we propose our new post-processing method to further modify the predicted object bounding box obtained by DPM detector, so as to cover the object more precisely.

2.1. Initialization of object bounding box estimation

Given an input image I (shown in Fig.1(a)), we first compute a set of N windows $\mathcal{W} = \{w_1, \cdots, w_k, \cdots, w_N\}$ with corresponding Bayesian posterior probabilities, denoted as $\mathcal{S} = \{s_1, \cdots, s_k, \cdots, s_N\}$ (shown in Fig.1 (b)) using the objectness approach [14]. We set $N = 1000$, which ensures covering most objects even in very difficult images [14]. Based on the fact that the objectness is designed to capture all possible objects within an image, we assume it has the reliability for providing at least *one* good candidate window w^* which covers the object of interest. However, the window with the highest objectness score $\max(\mathcal{S})$ is not always an effective choice[15], which usually encompasses other noisy objects, or locates poorly on object target.

To extract a reliable window from the pool of 1000 windows, we design a recursive selective scheme shown in Fig.1 (c)-(g). Inspired by the success of visual saliency applied in object recognition, we compute the reference region $\mathcal{T}$ (shown in Fig.1 (d)) by thresholding the saliency map $\mathcal{M}$(shown in Fig.1 (c)). The value of saliency map $\mathcal{M}$ at pixel $I(i,j)$ is obtained by summing up the objectness scores of the windows that cover this pixel:

$$\mathcal{M}(i,j) = \sum_{k=1}^{1000} \mathcal{M}_k(i,j) \tag{1}$$

where,

$$\mathcal{M}_k(i,j) = \begin{cases} s_k, & \text{if } I(i,j) \in w_k, \forall w_k \in \mathcal{W}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Meanwhile, we also adaptively select windows with high score as candidates, according to the histogram of 1000 sorted windows (shown in Fig.1(e)). To avoid near duplicate candidate windows, we further perform non-maximum suppression (NMS) to get a finer set of candidates. Contrary to the common practice, which starts the suppression procedure from highest scoring windows, we randomly choose one, for the reason that the highest scoring window is not necessarily the best. Fig.1 (f) illustrates the derived smaller set of n confident candidates $\hat{\mathcal{W}} = \{\hat{w}_1, \cdots, \hat{w}_i, \cdots, \hat{w}_n\}$, and their corresponding score denoted as $\hat{\mathcal{S}} = \{\hat{s}_1, \cdots, \hat{s}_i, \cdots, \hat{s}_n\}$.

Given the reference region $\mathcal{T}$ which implies the most salient region within an image, and confident candidate windows, the overlap between them provides valuable information to find the location of target object. The final estimate of the initial object bounding box w^* (Fig.1(g)) is determined

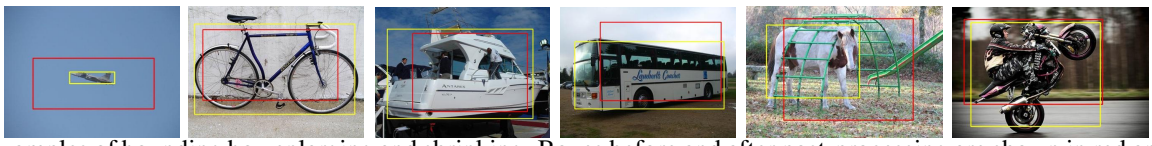

Fig. 2. Examples of bounding box enlarging and shrinking. Boxes before and after post-processing are shown in red and yellow, respectively.

by optimizing the following function:

$$w^* = \underset{\hat{w}_i \in \hat{\mathcal{W}}, \hat{s}_i \in \hat{\mathcal{S}}}{\arg\max} \quad \gamma \hat{s}_i + (1-\gamma)\frac{area(\mathcal{T} \cap \hat{w}_i)}{area(\mathcal{T} \cup \hat{w}_i)}, \quad i \in [1, n] \tag{3}$$

where γ is a parameter used to control the influence of the objectness score s_i. In practice, we set $\gamma = 0.2$.

2.2. Detection with deformable part-based models

We start training the DPM detectors with the derived bounding boxes from Section 2.1, which are treated as our positive training windows. Similarly to [2], each root filter hypothesis in a positive training image is initialized with the corresponding derived bounding box (ground-truth bounding box is used in [2]), and it is allowed to move around in a small neighborhood to maximize the filter score to compensate for imprecise bounding box estimation from Section 2.1. We refer the reader to [2] for more details concerning the DPM training and detection procedures. As in [5], we represent an image by a multiscale HOG feature pyramid [1] of 16 levels. For our DPM model, we use only a single component, since the multiple components are used for detecting objects with different views. We set the number of parts in DPM as 8 in all our experiments. And for negative training examples, we use random negatives from other object classes.

2.3. Bounding box post-processing

In many cases, the bounding boxes generated by DPM detectors are too large (resp. small) when detecting very small (resp. large) objects due to the restrictions of the size of the root filter and the scale of the feature pyramid. To improve the localization and to obtain a more precise estimate of the bounding box aspect ratio, we post-process each bounding box by enlarging or shrinking it to cover the object as much as possible. This is done using an improved version of the method proposed in [16] which measures the amount of area that the edge energy occupies. In brief, we first augment the original bounding box to 120% of the original width and height (*i.e.* 144% in total area), and calculate the absolute values of the gradients over the augmented bounding box and set the values which are less than 10% of the maximum to 0. To easily calculate the edge spatial distribution, then we resize the gradient magnitude image size to 100×100 and normalize the image sum to 1. Finally, we expand the bounding box in 4 directions from the centroid and stop until it contains 98% of the total gradient magnitude (edge energy) in the augmented box. This post-processing technique is not only able to crop out plain background regions, but also can expand to cover the foreground regions which are not encompassed by the original box. However, the cropping method in [5] is probably to fail with the latter. Fig. 2 shows a few examples of our bounding box post-processing results. It is also worth noticing that this post-processing technique works efficiently for the objects with a unique or plain background, but has limited help for those with cluttered or textured background.

3. EXPERIMENTAL EVALUATION

Dataset: Following the protocol of previous works [4, 5, 10], we evaluate the performance of our proposed weak supervision framework on two subsets from the training and validation set (*trainval*) of the PASCAL VOC 2007 dataset (*VOC07*)[13]: *VOC07-6×2* and *VOC07-14*. The *VOC07-6×2* subset contains 6 classes with *Left* and *Right* views (aspects) of each class, resulting in a total of 12 separating classes. The *VOC07-14* subset (same with *PASCAL07-all* defined in [5]) consists of 42 class/view combinations covering 14 classes and 5 views. Similar to [5], we remove all the images annotated as *difficult* or *truncated* in both training and evaluation steps.

Evaluation criteria: To make fair comparisons, we only choose the detection window with highest score per image, although our method can detect multiple instances appeared in the image using sliding window approach. We also report both results for initial and refined localization as [5, 10]. A refined localization is obtained by an iteratively trained DPM detector for one/several iteration(s) to refine the initial detection using the previous annotations as ground truth. Performance is evaluated with the percentage of training images in which an object is correctly covered by the window, if the strict PASCAL-overlap criterion is satisfied (intersection-over-union > 0.5).

Experimental evaluation: As Table 1 shows, our method outperforms [4] and our baseline approach [5] on both datasets. Our average performance of initial detection before cropping boxes on the *VOC07-6×2* and *VOC07-14* subsets is 38.74% and 21.73% respectively, versus 37.22% and 19.98% for [5]. These improvements are due to the initial object estimate of our method described in Section 2.1, which gives a better initialisation of the root filter of DPM detectors. We can also observe that both the cropping post-processing method from [5] (*i.e.* ours-[5] in Table 1) and our enlarging-or-shrinking (*i.e.* ours-ES) post-processing method steadily improve the average localization accuracy. In particular, our ES cropping method is superior to that of [5], as

Table 1. Average detection results (in %) compared with state-of-the-art competitors on the two variations of the PASCAL VOC 2007 datasets.

	VOC07-6×2					*VOC07-14*				
	no post-processing		with post-processing			no post-processing		with post-processing		
	[5]	ours	[5]	ours-[5]	ours-ES	[5]	ours	[5]	ours-[5]	ours-ES
Initialization	37.22	**38.74**	44.62	47.85	**48.59**	19.98	**21.73**	23.00	24.20	**25.12**
Refinement 1	51.63	55.85	53.11	56.78	58.02	25.11	27.46	26.38	28.21	28.94
Refinement 2	56.99	**59.82**	59.31	63.31	**63.91**	27.69	28.95	29.39	**32.87**	32.82
Refinement 3	59.32	—	61.05	—	—	**28.98**	—	30.31	—	—
Result from [4]	50.00					26.00				

our cropped bounding box is not only able to shrink to crop out the background regions, but also capable of enlarging to cover the whole foreground object resulted by incomplete coverage of the original window. An example is shown in the last row of Fig. 3, where the target object (motorbike) is only partially localized by the initial detector (shown in red rectangles in the middle and right images) for both [5] and our method. However, in the final detection (shown in yellow), our method is able to enlarge the bounding box to nearly include the whole object, while [5] tends to crop out both foreground and background regions. The middle rows in Table 1 indicate that localization accuracy can benefit from the refinement process. It is worth mentioning that with a better initialisation, our models converge to a steady level of performance after one less round of costly re-training (*i.e.* 2 iterations) than [5], and achieve slightly better results in the mean time. The detailed comparisons for our method with the state-of-the-arts on the *VOC07-6×2* dataset are listed in Table 2. The results show that our method outperforms [5] for most of the categories. Especially, our method achieves the state-of-the-art results in some classes where the target object possesses the most salient regions in that category (*e.g. aeroplane, bus, horse*). Interestingly, even without refinement process, the accuracy for our method with certain category (*e.g. aeroplane left*) is superior to the competitors with the time-consuming refinement procedure. Fig. 3 visually compares some of our results with those of [5].

Table 2. Class-level localisation accuracy (in %) for the *VOC07-6×2* dataset for our method vs. [4, 5, 10].

	Initialisation			Refined by detector		
	ours	[5]	[10]	ours	[5]	[4]
aero left	**65.1**	55.8	39.1	**69.7**	65.1	58.0
aero right	**64.1**	61.5	50.0	**84.6**	82.1	59.0
bike left	31.3	**31.3**	28.4	85.4	**87.5**	46.0
bike right	42.0	**44.0**	30.6	54.0	**68.0**	40.0
boat left	9.1	4.6	**15.1**	**13.6**	2.3	9.0
boat right	9.3	9.3	**20.7**	14.0	7.0	**16.0**
bus left	23.8	23.8	**31.0**	**42.9**	28.6	38.0
bus right	**65.2**	52.2	35.1	69.6	47.8	**74.0**
horse left	**64.6**	60.4	48.5	**87.5**	83.3	58.0
horse right	**73.9**	67.4	45.2	76.1	**80.4**	52.0
mbike left	**64.1**	48.7	46.3	87.2	**92.3**	67.0
mbike right	70.6	**76.5**	55.3	82.4	**88.2**	76.0
average	**48.6**	44.6	37.1	**63.9**	61.1	50.0

4. CONCLUSION

In this paper, we proposed a model enhancing the weakly supervised learning by emphasizing the importance of location and size of the initial class-specific root filter of deformable part model (DPM). We follow the general setup of [5] and introduce several substantial improvements to the weakly supervised DPM. The main contributions included new approaches based on objectness approach in generating the initial candidate window estimates. Furthermore we designed a flexible enlarging-and-shrinking post-processing procedure to modify the output bounding boxes of DPM, which can effectively further improve the final accuracy. Experimental results on the challenging PASCAL VOC 2007 database demonstrate that our proposed framework is efficient and competitive with the state-of-the-arts.

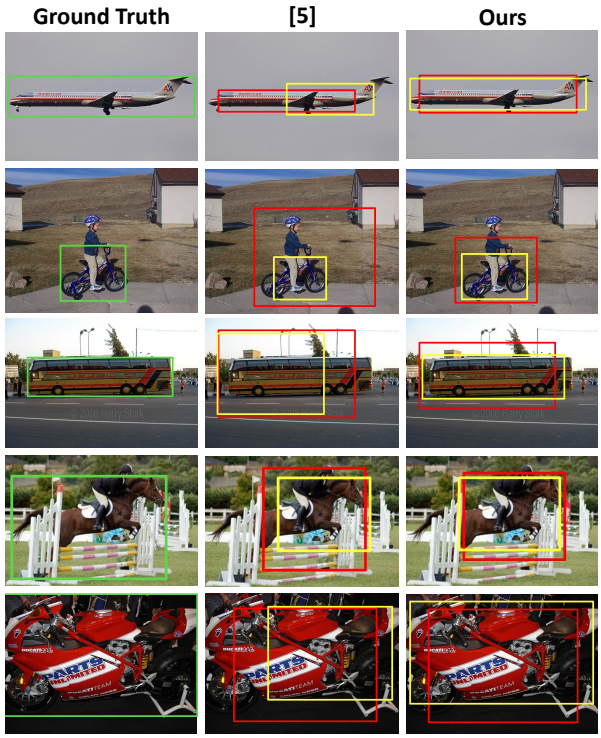

Fig. 3. Examples of detection results. The left column: ground-truth bounding boxes in green rectangles. The middle and right columns are detection results with [5] and our method, respectively. Initial detections are shown in red and detections refined by detectors are shown in yellow. Both results are with individual post-processing approach.

5. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR 2005*, 2005.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011.

[4] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *ECCV*, 2010.

[5] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.

[6] D. Crandall and D. Huttenlocher, "Weakly supervised learning of part-based spatial models for visual object recognition," in *ECCV*, 2006.

[7] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *ICCV*, 2009.

[8] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011.

[9] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *IJCV*, vol. 100, no. 3, pp. 275–293, 2012.

[10] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *ECCV*, 2012.

[11] R. Girshick, P. Felzenszwalb, and D. McAllester, "Object detection with grammar models," in *NIPS*, 2011.

[12] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *ECCV*, 2012.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, 2010.

[14] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *TPAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.

[15] Z. Shi, T. M. Hospedales, and T. Xiang, "Transfer learning by ranking for weakly supervised object annotation," in *BMVC*, 2012.

[16] X. Tang Y. Ke and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006.