Object-centric spatial pooling for image classification

Olga Russakovsky[1], Yuanqing Lin[2], Kai Yu[3], and Li Fei-Fei[1]

[1] Stanford University, {olga,feifeili}@cs.stanford.edu
[2] NEC Laboratories America ylin@nec-labs.com
[3] Baidu Inc. yukai@baidu.com

Abstract. Spatial pyramid matching (SPM) based pooling has been the dominant choice for state-of-art image classification systems. In contrast, we propose a novel *object-centric spatial pooling* (OCP) approach, following the intuition that knowing the location of the object of interest can be useful for image classification. OCP consists of two steps: (1) inferring the location of the objects, and (2) using the location information to pool foreground and background features separately to form the image-level representation. Step (1) is particularly challenging in a typical classification setting where precise object location annotations are not available during training. To address this challenge, we propose a framework that learns object detectors using only image-level class labels, or so-called weak labels. We validate our approach on the challenging PASCAL07 dataset. Our learned detectors are comparable in accuracy with state-of-the-art weakly supervised detection methods. More importantly, the resulting OCP approach significantly outperforms SPM-based pooling in image classification.

1 Introduction

Image object recognition has been a major research direction in computer vision. Its goal is two-fold: deciding *what* objects are in an image (classification) and *where* these objects are in the image (localization). Intuitively, if we know which objects are present, determining their location should be easier; alternatively, if we know where to look, recognizing the objects should be easier. Therefore, it is natural to think of these two tasks jointly [1–9].

However, in practice, classification and localization are often treated separately. Object localization is generally deemed as a harder problem than image classification even when precise object location annotations are available during training. In the purely image classification setting, it may be seen as a detour to attempt to localize objects. As a result, current state-of-the-art image classification systems don't go through the trouble of inferring object location information [10–14]. Most classification systems are based on spatial pyramid matching (SPM) [15] which pools low-level image features over pre-defined coarse spatial bins, with little effort to localize the objects [10–12].

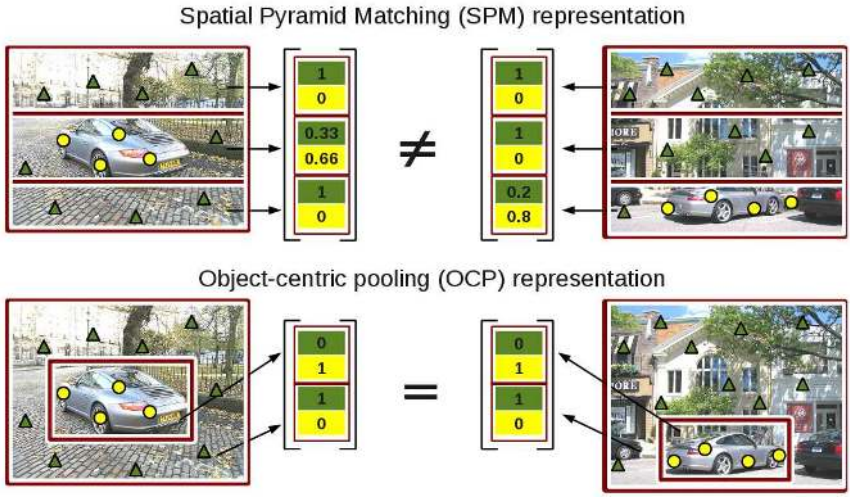

Fig. 1. We present *object-centric spatial pooling* (OCP), a method which first localizes the object of interest and then pools foreground object features separately from background features. In contrast, Spatial Pyramid Matching (SPM) based pooling [15] (top), the most common spatial pooling method for object classification, results in inconsistent image features when the object of interest (here, a car) appears in different locations within images, making is more difficult to learn an appearance model of the object. For the purpose of easy illustration, circles (yellow) denote object-related local features, triangles (green) denote background-related local features, and the numbers indicate the fraction of the respective local features in each pooling region.

This paper proposes a novel *object-centric spatial pooling* (OCP) approach for image classification. In contrast to SPM pooling, OCP first infers the location of the object of interest and then pools low level features separately in the foreground and background to form the image-level representation. As shown in Figure 1, if the location of the object of interest (a car in this case) is available, OCP tends to produce more consistent feature vectors than SPM pooling. Therefore, object location information can be very useful for further pushing the state-of-the-art performance of image classification.

Of course, the challenge for OCP is deriving accurate enough location information for improving classification performance. If the derived location information is not sufficiently accurate, it can end up hurting classification accuracy. There is interesting previous work on learning object detectors using only image-level class labels (or weak labels) [16, 17]. Although these methods yield impressive localization results, they are formulated as detection tasks and have not been shown to be helpful for improving image classification performance. Methods such as [1–7] attempt to localize objects to improve image classification accuracy but only demonstrate results on simple datasets such as subsets of Caltech101 classes. In contrast, we evaluate our proposed OCP method on the

highly cluttered PASCAL07 data [14], where we are able to localize objects with accuracy comparable to state-of-the-art weakly supervised object localization methods [16, 17] as well as to significantly improve image classification performance. To the best of our knowledge, this paper is the first to use weakly supervised object detection to improve image classification on PASCAL07, which is considered a challenging object detection dataset even when bounding box annotations are provided for training.

2 Related work

Classification. Many state-of-the-art image classification systems follow the popular image feature extraction procedure [10–12] shown in Figure 2. First, for each image, low-level descriptors like DHOG [18] or LBP [19] are sampled on a dense grid. They are then coded into higher dimensions through vector quantization, local coordinate coding (LCC) [10] or sparse coding [12]. Finally the coded vectors are pooled together, typically using SPM [15] pooling, to form the image-level representation. Much research in image classification has been focused on the former two steps, namely on different types of low-level descriptors [18–20] and coding methods [10, 12, 21–23]. In this paper we focus on the spatial pooling step, replacing the popular SPM with our object-centric pooling.

Methods such as [1–7] use localization information learned in a weakly supervised way to help boost classification accuracy by focusing on pooling low-level object features without background features. However, most of them only validate their approach on less cluttered and mostly centered datasets such as subsets of Caltech101 categories, Oxford Flowers 17 dataset, etc. For example, recently Feng et al. [7] presented a geometric pooling approach which resizes each image to the same size and learns a class-specific weighting factor for each grid position in an image. On the Caltech101 dataset, where most images are roughly aligned and centered, this method greatly improves over the previous state-of-the-art [10]. However, it has difficulty handling cluttered images like the ones of PASCAL07 [14]. Further, Nguyen et al. [1] and Bilen et al. [2] explicitly mention that some degree of context information (like road for cars) needs to be included into the detected object bounding box in order to be useful for image classification. This leads to very rough object localization even on simple datasets. In contrast, our work deals with high intra-class variability in object location and

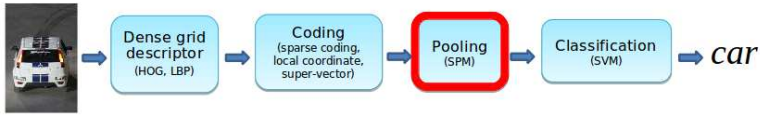

Fig. 2. A popular image classification pipeline of state-of-the-art methods [10–12]. In this paper we focus on the pooling step and propose an object-centric spatial pooling approach which achieves superior classification accuracy compared to the SPM pooling.

our proposed generic object-centric spatial pooling approach yields both classification improvements as well as competitive object localization results on the challenging PASCAL07 data.

If object location information is available during training, methods such as [24, 25] have been used to detect the object of interest, and [8, 9] showed how to use the output of object detectors to boost classification performance. There are two main differences compared our approach. First, we focus on the purely classification setting where no annotations beyond image-level class labels are available during training. Second, we learn a joint model for both localization and classification instead of combining the scores of the two tasks as post-processing.

Weakly supervised localization. There is a large body of work on weakly supervised object localization [16, 17, 26–28]. Most of these methods use HOG-type low-level features [18] which are faster for detection but have been shown to be inferior than bag-of-words models for classification [10, 25]. The current state of the art is the work of Pandey and Lazebnik [17] which uses deformable parts-based models [24] trained discriminatively in a weakly supervised fashion for object localization. In contrast, our goal here is image classification (not object localization) although we do utilize localization as an intermediate step.

3 Object-centric spatial pooling (OCP) for image classification

Let's first use an empirical experiment to quantitatively see how object location information can dramatically improve image classification performance. On the PASCAL07 classification dataset [14], we trained two classifiers for each object class: one classifier using features extracted from the full image, and the other classifier using features extracted only from the provided tight bounding boxes around the objects. We followed [10] in extracting image features and training linear classifiers. Both classifiers were trained on the training set and tested on the validation set. The former classifier (trained on full images) yielded 52.0% mean average precision (mAP), whereas the latter classifier (trained and tested on tight bounding boxes) achieved an astonishing 69.7% mAP. In comparison the current state-of-the-art classification result with a single type of low-level descriptor (which used a more involved coding method as well as significant post-processing) [11] is just 59.2% mAP. Therefore, it is evident that learning to properly localize the object in the image holds great promise for improving classification accuracy.

Now, the challenge is deriving accurate enough location information to help classification. Obviously, if the location information is not reliable enough, it can easily end up hurting classification performance instead. Reliable localization becomes very challenging on generic dataset like PASCAL07 [14] where objects vary greatly in appearance and viewpoint, are often occluded, and appear in highly cluttered and unstructured scenes. In fact, most work on weakly supervised localization uses simpler datasets [1, 2, 26–28]. Recently, Deselaers et al. [16] were the first to tackle PASCAL07. To simplify the problem, however,

they trained object class models separately for different viewpoints of objects. We are interested in learning generic object detectors without any additional annotations and evaluating classification performance on the original 20 object classes. To the best of our knowledge we are the first to do so.

To this end, we introduce a novel framework of *object-centric spatial pooling* (OCP) for image classification. OCP consists of two steps: (1) inferring the location of the objects of interested; and (2) pooling low-level features from the foreground and the background separately to form the image-level representation. In order to infer the object locations, we propose an iterative procedure for learning object detectors from only image class labels (or weak labels). Very different from existing methods for learning weakly supervised object detectors [16, 17], our approach directly optimizes the classification objective function and uses object detection as an intermediate step. This is described in Section 3.1. More importantly, OCP enables feature sharing between classification and detection: the resulting feature representation of OCP can be seen as both a bounding box representation (for detection) and an image representation (for classification). This is described in detail in Section 3.2. As we show in Section 4, such feature sharing plays an essential role in improving classification performance.

3.1 Classification formulation

We assume we are dealing with the binary image classification problem since multi-class classification is often solved in practice by training one-versus-all binary classifiers. Given N data pairs, $\{\mathbf{I}_i, y_i\}_{i=1}^N$, where I_i is the i^{th} image and $y_i \in \{+1, -1\}$ is a binary label of the image, the SVM formulation for binary image classification with OCP becomes

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i^N \xi_i \tag{1}$$

$$\text{s.t.} \quad y_i \max_{B\in\mathcal{BB}(i)} \left[\mathbf{w}^T\mathcal{P}_B(I_i) + b\right] \geq 1 - \xi_i \tag{2}$$

$$\xi_i \geq 0 \quad \forall i \tag{3}$$

where $\mathbf{w}$ is SVM weight vector, b is bias term, $\mathcal{P}_B(I_i)$ is the image feature representation of image I_i using OCP with given bounding box B, and $\mathcal{BB}(i)$ is the collection of all bounding box windows within image I_i. $\mathcal{BB}(i)$ can be obtained by either densely sampling sliding windows or by using salient regions [25]. We do not require any ground truth localization information in this optimization.

Interestingly, the above formulation can also be viewed as multi-instance learning (MIL) for object detection [1]. However, as in [1], the traditional MIL formation often only uses the foreground for constructing the bounding box features and discards the background information. This has its drawbacks in both detection and classification. As a result, the method of [1] was not able to accurately localize objects even on simpler datasets such as Caltech101; it tended to choose regions which were larger than the object of interest to encompass contextual information for classification. We fix these drawbacks by using

a foreground-background representation, as described below. As a result, we are able to localize objects on the significantly more challenging PASCAL07 [14] with accuracy comparable to state-of-the-art weakly supervised object localization methods [16, 17].

3.2 Foreground-background feature representation

In the classification formulation in Eq. 3, the foreground-background feature representation of OCP provides a natural mechanism for feature sharing between classification and detection. In fact, even for standalone detection and classification, the foreground-background feature representation is advantageous compared to traditional foreground-only feature representation.

Foreground-background for classification. The foreground-background feature representation provides stronger classification performance than its foreground-only counterpart. This is not surprising since the background provides strong scene context for classification [4, 29]. For example, for the class *boat*, the surrounding water in the image may provide a strong clue that this image contains a boat; similarly, seeing road at the bottom of an image can strongly indicate that this image is likely about *cars*. Going back to the classifiers trained on the tight bounding boxes as described at the beginning of Section 3, if we replace the foreground-only feature representation with the foreground-background representation, we further improve the classification mAP from 69.7% to 71.1%. This highlights the fact that the foreground-background feature representation carries important information for classification which may be missing in the foreground-only representation. This is illustrated in Figure 3.

Foreground-background for detection. Object detectors trained with the foreground-background features also tend to yield more accurate bounding boxes during detection. Since the foreground and background models are learned jointly, they will prevent the object appearance features from leaking into the background, and context features from leaking into the foreground. This is illustrated in Figure 4. To validate the effectiveness of the foreground-background feature representation for detection, we also experimented on PASCAL07, training fully supervised object detectors using the foreground-only and the foreground-background feature representation respectively. It was no surprise that the foreground-background feature representation yielded significantly better detection performance. Here we skip the details of the experiments for simplicity since supervised detection is not the major focus of this paper. In Figure 6 in the experimental

Fig. 3. Example images which were misclassified using just the foreground representation but correctly classified when using the foreground-background representation.

Fig. 4. Bounding boxes bb_1 and bb_2 have a similar foreground-only feature representation, but they are very different under the foreground-background representation. Here, the numbers denote the count of object-related descriptors. For bb_1, parts of object that leaked into the background will be greatly discounted by the background model.

results section, however, we show the differences in detections made with the foreground-only and the foreground-background model in our OCP framework.

With the foreground-background representation of OCP, optimizing the formulation in Eq. 3 can be seen as a simultaneous detection and classification procedure. This is because the foreground-background representation can be seen as both a bounding box representation (for detection) and an image-level representation (for classification).

3.3 Optimization

Now that we have defined our objective and our foreground-background feature representation, we discuss how to optimize this formulation. The optimization in Eq. 1 is non-convex because of the maximization operation in the constraints, thus we need to be careful during optimization to avoid local minima. In particular, since we are not given any localization information during training, our optimization algorithm consists of an outer loop that bootstraps the background region from the foreground and an inner loop that trains the apperance model.

Outer loop: bootstrapping background regions. In a purely classification setting, no foreground and background annotations are provided initially. We initialize the background region by cropping out a 16-pixel border of each image. Then the outer loops bootstraps the background by gradually shrinking the smallest bounding box considered in the bounding box search ($\mathcal{BB}(i)$ in Eq. 1). Thus we begin localizing using large windows and iteratively allow smaller and smaller windows as we learn more and more accurate models. As the background region is allowed to grow, the algorithm learns more and more accurate background models. If the algorithm goes too aggressively, it will end up in bad local minima. For example, if the localization is so inaccurate that many features from the object of interest appear in the background region, the model would learn that objects features actually belong to the background. This would lead to bad classification models which are hard to correct in later iterations. However, as long as such bad local minima are avoided, the specific rate of shrinking the foreground region does not affect performance in our experiments.

Inner loop: learning the appearance model for detection. Given the current constraint on the background size, we need to learn the best object ap-

pearance model. This is done in two steps: (1) detection, where given the current appearance model we find the best possible object location from positive images (images that are known to contain the object of interest); and (2) classification, where given the proposed bounding boxes from positive images as positive examples and a large sample of bounding boxes from negative images as negative examples, we construct the bounding box representation using OCP and then train a binary SVM classifier for discriminating the positive bounding boxes from the negative bounding boxes. In contrast to more common treatments which would need another loop to bootstrap the difficult negative bounding boxes and iteratively improve the SVM model, here we get rid of this loop by solving an SVM optimization directly with all (often millions) negative bounding boxes.

We make use of the candidate image regions proposed in an unsupervised fashion by [25] to avoid both sampling too many negative windows for classification and running sliding windows search for detection. Since the candidate bounding boxes aim to achieve high recall rate ($> 96\%$), we ended up with 1000$\sim$3000 candidate bounding boxes per image. For PASCAL07, we have 5011 images in the training and validation sets. Therefore, for each inner loop, we need to solve for 20 binary SVMs with about 10 million data examples. Furthermore, our feature representation for OCP is very high-dimensional: we used a codebook of 8192 for LLC coding [10], pool the low-level features on the foreground region using 1×1 and 3×3 SPM pooling regions [15], and separately pool all low-level features features in the background, thus resulting in a feature vector of dimension $8192 \times 11 = 90112$. Indeed, if we save all the feature vectors from the 5011 images, this would require more than 700G of space. Most off-the-shelf SVM solvers would not be able to handle such a large-scale problem. So, we developed a stochastic gradient descent algorithm with averaging using a similar idea to [30]. We were able to run an inner loop in 7$\sim$8 hours and to finish the training (inner look and outer loop) in about 3 days on a single machine.

4 Experiments

We validate our approach on the challenging PASCAL07 dataset [14], containing 5011 images for training and validation, and 4952 images for testing. This dataset consists of 20 object categories, with object instances ocurring in a variety of scales, locations and viewpoints.

Image representation. For low-level features, we extract DHOG [18] features with patch sizes 16×16, 25×25, 31×31 and 46×46. We then run Linear Locality-Constrained (LLC) coding [10] using a codebook of size 8192 and 5 nearest neighbors. For the baseline representation, we pool the DHOG features using 1×1 and 3×3 SPM pooling regions [15] over the full image. Thus each image is represented using a feature vector of dimension $8192 \times 10 = 81920$. For our object-centric pooling, we use the same SPM representation but on the foreground region and also pool over all low-level features in the background separately, thus giving us a feature dimension of $8192 \times 11 = 90112$.

4.1 Joint classification and localization

The main insight behind our approach is that object classification and detection can be mutually beneficial. In particular, as the classification accuracy improves we expect detection accuracy to improve as well, and vice versa. We begin by verifying that this is indeed the case. Figure 5 shows the steady improvement in mean average precision on both classification and detection over the iterations (outer loop) of our algorithms. As a baseline (iteration 0), we use a classifier trained on full images with the SPM spatial pooling representation, which is equivalent to assuming an empty background region in foreground-background representation. Interestingly, even after just one iteration, our classification mAP is already 54.8%, which is 0.5% greater than the 54.3% SPM classification result.[1] In the end our OCP method achieves 57.2% classification mAP, significantly outperforming the SPM representation. In fact, it significantly outperforms even a much richer 4-level SPM representation of size 8192×30 which achieves only 54.8% classification mAP. On the detection side, our approach was able to improve the baseline of 6.10% detection mAP to the final 15.0%.

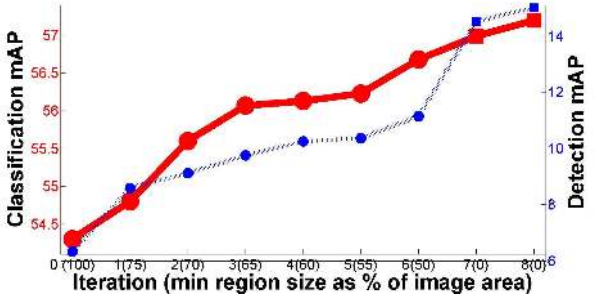

Fig. 5. Classification and detection mAP on the PASCAL07 test set over the iterations of our joint detection and classification approach. The red solid line is classification mAP, and the blue dotted line is detection mAP. We see a steady joint improvement of classification and detection accuracy.

It is important to note that jointly optimizing detection and classification using OCP as in Eq. 3 plays an essential role in achieving the joint improvements for classification and detection. As we show below, when detection and classification are optimized separately, higher detection accuracy may not always means higher classification accuracy.

[1] We make use of only one type of low-level image descriptor in contrast to [9, 31], and don't do any additional post-processing of the features in contrast to [10, 11]. The work of [10] gives 59.3% classification mAP on this dataset when using LLC coding, but this relied on significant post-processing of the resulting image features. To simplify the comparison, we do not involve the post-processing.

4.2 Image classification

OCP significantly boost of classification accuracy on most of the 20 object classes, as shown in Table 1. In particular, OCP achieves significant improvement on the following categories: dog (7.3% improvement), bottle (7.1%), bicycle (6.8%), sheep (6.2%), diningtable (5.9%), bus (4.6%), motorbike (4.3%) and even 1.3% on the notoriously difficult potted plant category. Noticeably, many of these categories are relatively small objects (like bottles) embedded in cluttered environments. OCP greatly improves classification accuracy on these categories by making an effort to localize the objects.

Method	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining
SPM	72.5	56.3	**49.5**	63.5	22.4	60.1	76.4	57.5	**51.9**	42.2	48.9
OCP	**74.2**	**63.1**	45.1	**65.9**	**29.5**	**64.7**	**79.2**	**61.4**	51.0	**45.0**	**54.8**

Method	dog	horse	motbike	person	plant	sheep	sofa	train	tv	Mean
SPM	38.1	75.1	62.8	82.9	20.5	38.1	46.0	**71.7**	50.5	54.3
OCP	**45.4**	**76.3**	**67.1**	**84.4**	**21.8**	**44.3**	**48.8**	70.7	**51.7**	**57.2**

Table 1. Classification AP of object-centric spatial pooling compared to the standard SPM spatial pooling on the PASCAL07 test set.

There are three categories that proved difficult for OCP to improve: chairs (-0.9%), trains (-1.0%) and birds (-4.4%). For the bird and chair categories, the objects are often occluded (e.g., birds are often occluded by trees, and chairs are often occluded by people sitting on them), which make them very challenging for detection even when bounding box annotations are available (see [24, 14]). For the slight drop in the train category, since trains are already relatively well-centered in images, SPM pooling alone yields very satisfactory classification accuracy (71.7%) and is difficult to further improve.

We also investigate using the foreground-only (instead of the foreground-background) feature representation when optimizing Eq. 3.[2] This foreground-only representation leads to an improvement from the baseline SPM model – the mAP increases from 54.3% to 55.7%. This is a 1.4% improvement as compared to the 2.9% improvement as in the case of our foreground-background representation. Figure 6 illustrates some location results, showing that foreground-background representation often yields better localization.

4.3 Weakly supervised object localization

Even though our primary goal is image classification, the proposed object-centric spatial pooling also accurately localizes the objects of interest. PASCAL07 is

[2] This experiment is a more assertive version of the technique described in Nguyen et al. [1]: the optimization framework is similar to [1] but with significantly stronger low-level descriptors (HOG descriptors [18] with LLC coding [10] compared to vector-quantized SIFT [20]) and with much more negative training data.

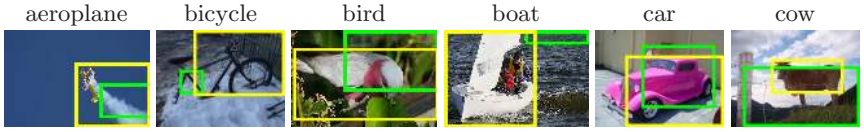

Fig. 6. Images where object-centric pooling with the foreground-background model (yellow) localizes objects more accurately than the foreground-only model (green).

a very challenging dataset for weakly supervised localization (where bounding box information is not available during training). Only a few recent works have tackled this data (Deselaers et al. [16] and Pandey and Lazebnik [17]). They focused on localizing only a handful of the object classes and use the available viewpoint annotations during training to assist learning. In contrast, we work on the full dataset without using these additional annotations to mimic the purely classification setting.

Weakly supervised localization can be evaluated directly on the training set (in our case the PASCAL07 trainval set) since only image-level class labels are available during training. Following [16, 17] we compute localization accuracy as the percentage of training image in which an instance was correctly localized by the highest-scoring detection according to the PASCAL criterion (window intersection over the union $\geq 50\%$). On the 14 classes of PASCAL07-*all*[3] introduced by [16], our localization accuracy is 27.4%, which is comparable to 26% of [16] using additional viewpoint annotations and 30.0% of [17].

As we're most interested in inferring object location on unseen images, we evaluate the detection accuracy on the test set as well. Table 2 compares our detection average precision on six PASCAL07-6x2 classes [16] evaluated on all test images with the current state-of-the-art in weakly supervised localization. We obtain 22.8%, outperforming the previous best 20.8% of [17] which used additional viewpoint annotations. On all 20 classes, we obtained 15.0% detection mAP compared to 29.1% mAP of the state-of-the-art deformable part-based model that used bounding box labels for detector training [24].

Method	aeroplane		bicycle		boat		bus		horse		motorbike		Average
	left	right	left	right	left	right	left	right	left	right	left	right	
Deselaers [16]	9.1	23.6	33.4	49.4	0.0	0.0	0.0	16.4	9.6	9.1	20.9	16.1	16.0
Pandey [17]	7.5	21.1	38.5	44.8	0.3	0.5	0	0.3	45.9	17.3	43.8	27.2	20.8
OCP	**30.8**		25.0		**3.6**		**26.0**		21.3		29.9		**22.8**

Table 2. Comparison of detection AP on the PASCAL07-6x2 test set for our method versus [16, 17]. Both [16, 17] split up the objects by left and right viewpoint to make the models easier to learn. We do not make use of these additional labels and learn a single model for each object.

[3] PASCAL07-*all* includes all classes of PASCAL07 except bird, car, cat, cow, dog and sheep. [16]

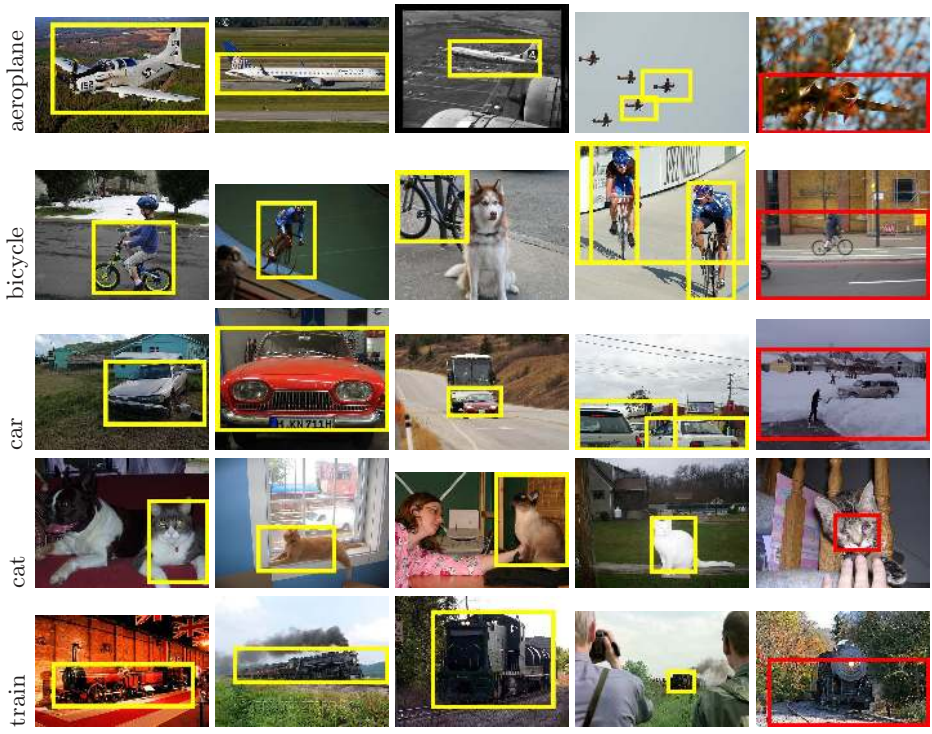

Fig. 7. Foreground regions detected by the object-centric pooling framework on PASCAL07 test images. The models are learned without any ground truth localization information. Yellow boxes correspond to correct detections and red boxes are failed detections. On images where multiple instances of a object class are presented, we show the top few detections after running non-maximal suppression.

Figure 7 shows some examples of our detection results on PASCAL07 test set. Localization is often quite reasonable, which is amazing considering the difficulty of the dataset and the lack of any bounding box annotations during training. Even on images with multiple object instances our method is sometimes able to separate out the different instances.

Interestingly, when we used the location information derived from the deformable part-based model mentioned above [24] learned with the help of bounding box annotations, images features constructed using our image representation with the foreground-background pooling yielded a classification mAP of 56.9%. This is inferior to the aforementioned 57.2% classification mAP obtained using OCP, where our proposed approach in Eq. 3 did not use any bounding box annotations and only achieved 15.0% detection mAP. This strongly highlights the importance of the formulation in Eq. 3, which uses classification as the major optimization objective and jointly optimizes detection and classification when solving the optimization.

5 Conclusion

We presented an object-centric spatial pooling (OCP) approach for improving classification performance. The challenge of OCP is training reliable object detectors with no available bounding box annotations as in a typical classification setting. We propose a framework that directly optimizes classification objective with detection being treated as an intermediate step. The key to this framework is the foreground-background feature representation by OCP that naturally enables feature sharing between detection and classification. Our results on the challenging PASCAL07 dataset show that not only is the proposed OCP approach able to improve the classification accuracy compared to using SPM pooling, but it also yields very reasonable object detection results. We believe this is an important step toward better image understanding – not only deciding *what* objects are in an image but also figuring out *where* these objects are.

Our future work includes incorporating bounding box annotations during training (from all or just a subset of images) to further improve the classification performance. We are also very interested in exploiting even more powerful visual features than the simple LLC feature as used in this paper. As demonstrates by the motivation experiment described in the beginning of Section 3, there is much room for improving classification performance by utilizing location information. This paper is just an initial step toward that direction.

Acknowledgements

This work was done while Olga Russakovsky was a summer intern and Kai Yu was a research staff member at NEC Labs. Li Fei-Fei was supported partially by a MURI grant from ONR. Many thanks to Jia Deng at Stanford and to Anelia Angelova, Timothee Cour, Chang Huang and Shenghuo Zhu at NEC Labs for helpful discussions.

References

1. Nguyen, M.H., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV. (2009)
2. Bilen, H., Namboodiri, V.P., Gool, L.V.: Object and action classification with latent variables. In: BMVC. (2010)
3. Chai, Y., Lempitsky, V., Zisserman, A.: BiCoS: A bi-level co-segmentation method for image classification. In: CVPR. (2011)
4. K.Murphy, Torralba, A., Eaton, D., Freeman, W.: Object detection and localization using local and global features. Lecture Notes in Compute Science (2006)
5. Crandall, D., Huttenlocher, D.: Weakly supervised learning of part-based spatial models for visual object recognition. In: ECCV. (2006)
6. Zhang, Y., Chen, T.: Weakly supervised object recognition and localization with invariant high order features. In: BMVC. (2010)
7. Feng, J., Ni, B., Tian, Q., Yan, S.: Geometric ℓ_p-norm feature pooling for image classification. In: CVPR. (2011)

8. Hedi, H., Frederic, J., Cordelia, S.: Combining efficient object localization and image classification. In: ICCV. (2009)
9. Song, Z., Chen, Q., Huang, Z., Hua, Y., Yan, S.: Contextualizing object detection and classification. In: CVPR. (2011)
10. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained Linear Coding for image classification. In: CVPR. (2010)
11. Zhou, X., Yu, K., Zhang, T., Huang, T.: Image classification using super-vector coding of local image descriptors. In: ECCV. (2010)
12. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR. (2009)
13. Berg, A., Deng, J., Satheesh, S., Su, H., Fei-Fei, L.: Large scale visual recognition challenge. http://www.image-net.org/challenges/LSVRC/2011/ (2010-2011)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) challenge. IJCV **88** (2010) 303–338
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial Pyramid Matching for recognizing natural scene categories. In: CVPR. (2006)
16. Deselaers, T., Alexe, B., Ferrari, V.: Localizing objects while learning their appearance. In: ECCV. (2010)
17. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV. (2011)
18. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR. (2005)
19. Ahonen, T., Hadid, A., Pietikinen, M.: Face description with local binary patterns: Application to face recognition. PAMI **28** (2006)
20. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
21. Huang, Y., Huang, K., Tan, T.: Salient coding for image classification. In: CVPR. (2011)
22. Gao, S., Chia, L.T., Tsang, I.W.: Multi-layer group sparse coding – for concurrent image classification and annotation. In: CVPR. (2011)
23. Perronnin, F., Sanchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV. (2010)
24. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI **32** (2010)
25. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In: ICCV. (2011)
26. Russell, B.C., Freeman, W.T., Effros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR. (2006)
27. Kim, G., Torralba, A.: Unsupervised detection of regions of interest using iterative link analysis. In: NIPS. (2009)
28. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: CVPR. (2007)
29. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences **11** (2007)
30. Lin, Y., Lv, F., Cao, L., Zhu, S., Yang, M., Cour, T., Yu, K., Huang, T.: Large-scale image classification: Fast feature extraction and SVM training. In: CVPR. (2011)
31. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classification. In: CVPR. (2010)