Adaptive Region Pooling for Object Detection

Yi-Hsuan Tsai[1], Onur C. Hamsici[2] Ming-Hsuan Yang[1]
[1]UC Merced. [2]Qualcomm Research, San Diego.

Learning models for object detection is a challenging problem due to the large intra-class variability of objects in appearance, viewpoints, and rigidity. We address this variability by a novel feature pooling method that is adaptive to segmented regions. The proposed detection algorithm automatically discovers a diverse set of exemplars and their distinctive parts which are used to encode the region structure by the proposed feature pooling method. Based on each exemplar and its parts, a regression model is learned with samples selected by a coarse region matching scheme.

Adaptive Region Pooling. We propose to find a diverse set of exemplars that represent the variations in the training set. Training samples are grouped according to their similarity of region proposals and representative exemplars are selected from each of these groups. In the training phase, we use each of these exemplars to search for training samples that have similar regions.

For each representative exemplar found in the training set, we aim to discover parts within the object bounding box based on the segmentation. Unlike the conventional approaches that define the parts as a set of rectangular regions, we present a method that allows to precisely extract non-rigid deformable regions. We apply several rules that determine if a segment can be an object part. Details and examples for parts are described in the paper.

We also define our feature pooling algorithm according to the parts of each exemplar in the previous steps. Unlike spatial pyramid pooling that is defined over a pre-defined grid, our pooling method aims to match meaningful segments from the exemplar to the target regions. We illustrate the procedure in Figure 1.

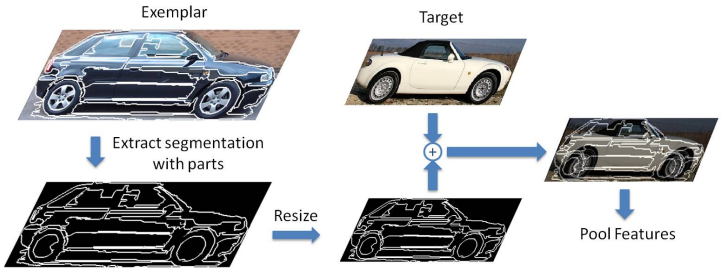


Figure 1: Our feature pooling procedure. Given an exemplar with parts, we resize the region structure to the same size as the target region. The resized part mask is then applied to the target region for pooling features on each part.

Multiple Exemplar-based Models. In this work, we learn a linear SVR model for each representative exemplar. A set of training samples that are similar to the exemplar are obtained by a coarse region matching procedure. Given an exemplar with the object mask M^e, which is the union regions of parts, we compute the similarity score between M^e and a target region R based on the appearance and the size of the region:

$$S(M^e,R) = \langle \mathbf{z}^e, \mathbf{z}^r \rangle \cdot \left(\frac{\min(|M^e|,|R|)}{\max(|M^e|,|R|)} \right),$$

where $\mathbf{z}^e, \mathbf{z}^r$ are feature vectors, and $|M^e|$ and $|R|$ denote the size of an exemplar mask and a target region, respectively. We use the same coarse region matching scheme in the training and testing stage to ensure consistency in the sample space. In training, the coarse region matching allows us to select samples that are similar to one exemplar and enables us to learn a discriminative linear model. In testing, it eliminates a large set of easy negatives.

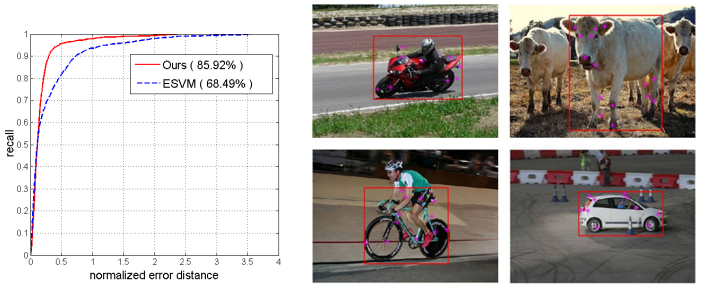
To learn a SVR model, we use the top N samples by coarse region matching in each positive image to learn a initial model. To refine the model, we run one iteration for negative mining by adding samples with regression scores larger than 0.3 among the top N samples in negative images. Note that the regression score is computed based on the union-over-intersection overlap between the bounding box of the ground truth annotation and region proposals.

Experimental Results. We conduct experiments on the object detection task of PASCAL VOC 2007 dataset. First, we compare the proposed algorithm with other exemplar-based methods [1, 2]. Second, we show that our approach accommodates the convolutional neural networks (CNN) features [3, 4] to achieve state-of-the-art results. Finally, we evaluate the performance of transferred object information to the detected objects using the proposed algorithm quantitatively and qualitatively.

Table 1: Detection mAP on the PASCAL VOC 2007 test set.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	dtable
ESVM [2]	20.4	**40.7**	9.3	10	10.3	31	40.1	9.6	**10.4**	**14.7**	2.3
LDA [1]	17.4	35.5	**9.7**	**10.9**	**15.4**	17.2	**40.3**	10.6	10.3	14.3	4.1
Ours	**31.5**	37.7	5.5	7.9	5	**33.5**	37.3	**32**	5	13.8	**27.2**
Regionlets (CNN) [4]	44.6	55.6	24.7	23.5	6.3	49.4	51	**57.5**	14.3	35.9	45.9
DPM (CNN) [3]	39.7	59.5	**35.8**	24.8	**35.5**	53.7	48.6	46	**29.2**	36.8	45.5
Ours (CNN)	**58.1**	**60.6**	31	**29.3**	17.8	**61**	**56.1**	55.9	18.1	**42.3**	**52.9**

	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
ESVM [2]	9.7	38.4	**32**	19.2	**9.6**	**16.7**	11	29.1	31.5	19.8
LDA [1]	1.8	**39.7**	26	**23.1**	4.9	14.1	8.7	22.1	15.2	17.1
Ours	**15.4**	25.6	31.7	13.8	1.3	16.2	**28.3**	**34**	**31.7**	**21.7**
Regionlets (CNN) [4]	41.3	**61.9**	54.7	**44.1**	16	28.6	41.7	**63.2**	44.2	40.2
DPM (CNN) [3]	42	57.7	56	37.4	**30.1**	31.1	**50.4**	56.1	51.6	43.4
Ours (CNN)	**46.9**	52	**58**	32.7	20.3	**43.7**	46.6	53.2	**57.6**	**44.7**



(a) (b)

Figure 2: Keypoints transfer results on the PASCAL VOC 2007 dataset. (a) shows the recall-error curve comparing to the ESVM method. The number in the legend indicates the recall rate when the error distance is 0.25. (b) visualizes transfered keypoints (marked in pink) results on detected objects.

[1] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.

[2] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[3] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos. Deformable part models with cnn features. In *ECCV workshop*, 2014.

[4] W. Y. Zou, X. Wang, M. Sun, and Y. Lin. Generic object detection with dense neural patterns and regionlets. In *BMVC*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.