IEICE TRANS. INF. & SYST., VOL.E98–D, NO.7 JULY 2015

LETTER

Discriminative Semantic Parts Learning for Object Detection

Yurui XIE[†a)], ***Nonmember***, **Qingbo WU**[†], ***Student Member***, ***and*** **Bing LUO**[†], ***Nonmember***

SUMMARY In this letter, we propose a new semantic parts learning approach to address the object detection problem with only the bounding boxes of object category labels. Our main observation is that even though the appearance and arrangement of object parts might have variations across the instances of different object categories, the constituent parts still maintain geometric consistency. Specifically, we propose a discriminative clustering method with sparse representation refinement to discover the mid-level semantic part set automatically. Then each semantic part detector is learned by the linear SVM in a one-vs-all manner. Finally, we utilize the learned part detectors to score the test image and integrate all the response maps of part detectors to obtain the detection result. The learned class-generic part detectors have the ability to capture the objects across different categories. Experimental results show that the performance of our approach can outperform some recent competing methods.

key words: *object detection, sparse representation*

1. Introduction

Object detection is one of the most challenging problem due to the variations in object appearance, pose, illumination and viewpoint, etc. Recently, the part-based models have received increasing research attention in computer vision [1]–[4]. They model an object as a set of important parts and achieve good performances for the object-detection problem. The deformable part-based model (DPM) method [1]–[3] represents an object using the deformable parts, and each part of the object describes the local appearance properties of an object. In order to further improve the performance, the strongly supervised information [2], [4] provided by the human-annotated is also incorporated into the learning processing of part detectors. Another direction aims to generalize the deformable part model from image to spatiotemporal data [3]. However, the main limitation for the above class-specific object detection approaches is that the learned model can only be applied to one specific object class, which is insufficient to model a new object category. In order to overcome the limitation, some other works [5]–[9] have attempted to measure the objectness of image patch with the help of saliency-based and segmentation-based cues. They can also achieve good performances for the class-generic object detection.

Inspired by these part-based methods for object detection, we propose a new detection framework by learning the object semantic parts with only the bounding box labels. The key observation is that even though the parts of different object categories might have variations in terms of the appearance and arrangement, the constituent parts still maintain the consistency in geometric properties. Compared with some existing part-based detection methods, the proposed approach has the following distinctive characteristics: First, the discriminative semantic part detectors can be automatically learned from bounding box labels by the proposed approach, without requiring heavy supervision or detailed label information. Second, different from some established part-based methods that can only detect the object of specific category using the trained object model, the learned class-generic part detectors have the ability to capture the objects across different categories. Figure 1 illustrates the flowchart of proposed approach for object detection.

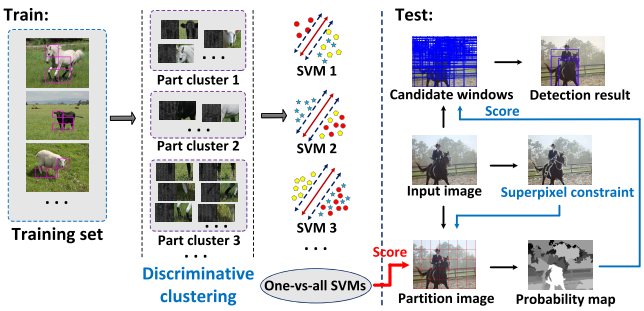

Fig. 1 The procedures of proposed object detection method.

2. Learning Discriminative Semantic Part Detectors

In this section, we introduce the details that how to generate these discriminative semantic part set of objects with only bounding box labels. In particular, the object parts should meet the following two main requirements: 1) Discriminability: the semantic parts should capture the local appearance properties of objects, which are helpful for detecting the object categories. 2) Repeatability: these learned parts should be repeated in the object classes constantly.

Given a large set of training images with only the object category labels $T = \{(I_i, B_i)\}_{i=1}^{N}$, where I_i is the i-th training image and B_i denotes the set of bounding boxes that describe the location of each object, N is the number of training images, our goal is to discover a set of discriminative semantic parts S from the training set T. Then these obtained part set S can be used to learn the object part detectors.

Manuscript received January 15, 2015.
Manuscript revised March 28, 2015.
Manuscript publicized April 15, 2015.

[†]The authors are with the University of Electronic Science and Technology of China, China.
a) E-mail: gloriousxyr@163.com

DOI: 10.1587/transinf.2015EDL8014

2.1 Discovering Semantic Parts

In order to obtain a set of initial object parts, we first regularly partition each object region by the bounding box from all the training images. Then we perform the unsupervised clustering to obtain the initial object part set $\Phi = \{C_q\}_{q=1}^{Q}$, where C_q denotes the q-th initial part cluster and Q is the number of initial part clusters. Specifically, the affinity propagation(AP) clustering [10] is applied on all the partitioned object parts. Unlike the traditional clustering method such as k-means, the procedure of affinity propagation clustering is done automatically, without requiring the parameter that specifies the desired number of clusters. This characteristic ensures that the number of part clusters is determined by the appearance properties of object parts, rather than the value is to be set manually.

In our method, each element of adjacent matrix in the AP clustering is measured by: $A(i, j) = d(f_i, f_j)$, where $A(i, j)$ denotes the similarity between the partitioned parts i and j, $d(\cdot)$ denotes the χ^2 distance metric function, f_i and f_j are the HOG descriptors [11] for the parts i and j, respectively. Then the similarities of all the pair of parts are computed accordingly and we can obtain the whole adjacent matrix A. This adjacent matrix is further fed into the AP clustering to obtain the initial object part set. To enforce the repeatability of object parts, the candidate part clusters are generated by removing these initial clusters with less than L members. In practice, we set this parameter $L = 10$. For the discriminability, we propose a refinement approach to rank all the candidate part clusters and select these discriminative part clusters in the following step.

2.2 Ranking Part Clusters By Sparse Representation

Given the set of candidate object parts that can be described as $\Psi = \{C_j\}_{j=1}^{M}$, where C_j denotes the j-th candidate part cluster and $M(< Q)$ is the size of candidate part set, we propose a refinement approach based on sparse representation to select these discriminative part clusters. Recently, the sparse representation [12] has attracted much attentions in computer vision, which has the ability to approximate the target signal as a linear combinations of a small number of atoms by a given dictionary. In our approach, each candidate part cluster is taken as the target data and is reconstructed by the rest of candidate part clusters. Mathematically, the representation coefficients for one part cluster C_j are computed by minimizing the following objective function:

$$\min_{A_j} \| X_j - D_{\bar{j}} \cdot A_j \|_2^2 + \lambda \| A_j \|_1 \tag{1}$$

where $X_j = [x_j^1, x_j^2, \ldots, x_j^n]$ denotes the target part data and each column $x_j^t, (t = 1, 2, \ldots, n)$ of matrix X_j is the HOG feature vector associated with one of the object part within the cluster C_j, n is the number of object parts in this cluster. $D_{\bar{j}}$ is the reconstruction dictionary that is built by all the feature vectors of the rest of candidate part clusters, A_j is the sparse representation coefficients of the j-th cluster over $D_{\bar{j}}$. The parameter λ is used to balance the different terms in the objective function. In our method, the algorithm [13] is adopted to solve the above optimization formulation.

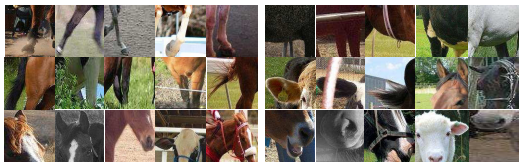

Fig. 2 Semantic part clusters: some examples within the clusters after refinement. The discovered object parts have the similarity appearance properties in each cluster and are inconsistent with the parts of other clusters in appearance space.

For each candidate part cluster, the construction residual is computed by $\| X_j - D_{\bar{j}} \cdot A_j \|_2^2$. Since the reconstruction residual measures the similarity between the current cluster C_j and the rest of candidate part clusters, we can use it to evaluate the discriminative power of current part cluster. The higher value of reconstruction residual, the more discriminative power of the current part cluster C_j. Finally, we can rank all the candidate part clusters by the values of construction residuals and the discriminative part clusters are selected to further train the part detectors. In the experiment, we select the top $K(< M)$ candidate part clusters that have the highest residual values to learn the discriminative detectors. Figure 2 shows some instances of visualization, each row contains two selected clusters and we demonstrate five object part instances within each cluster. Due to our discriminative clustering with sparse representation refinement, it can be observed that object parts within each cluster have the similarity appearance properties and are inconsistent with the parts of other clusters in appearance space.

2.3 Learning Discriminative Part Detectors

Once the K discriminative part clusters are obtained, we can use them to learn the semantic part detectors that have the ability to capture the local object regions in appearance space. For each discriminative part cluster, we train a linear SVM classifier in the one-vs-all manner. Specifically, we treat the object parts in current part cluster as the positive samples, and all the object parts from the rest of $K - 1$ part clusters as the negative samples to train SVM classifier corresponding to the current part cluster. Finally, the learned classifier is used as a part detector to capture these patches within an image that are similar to the parts of discriminative cluster. The learned semantic part detectors can be described as: $S = \{(C_n, V_n)\}_{n=1}^{K}$, where C_n is the n-th discriminative part cluster and V_n denotes the trained part detector associated with the n-th cluster, K is the number of learned part detectors. The whole approach of semantic part detectors learning is summarized in Algorithm 1.

3. Detecting Algorithm

In the test stage, we first partition the image into regular

Algorithm 1 Learning discriminative part detectors

Input: Training set $T = \{(I_i, B_i)\}_{i=1}^{N}$, parameters L, K and λ.
Output: Discriminative part detectors $S = \{(C_n, V_n)\}_{n=1}^{K}$

$\Phi = \{C_q\}_{q=1}^{Q} \Longleftarrow Construct\ (T)$ ◁ Construct the initial part set Φ with the affinity propagation clustering [10]
$\Psi = \{C_j\}_{j=1}^{M} \Longleftarrow Remove\ (\Phi)$ ◁ Generate the candidate part clusters Ψ by removing these clusters with less than L members
for $j = 1$ to M **do**
  $X_j \Longleftarrow ExtractHOG\ (C_j)$ ◁ Extract the HOG descriptors for all the members within the j-th candidate part cluster
  $D_{\bar{j}} \Longleftarrow BulidDictoinary\ (\{C_t\}_{t=1,t\neq j}^{M})$ ◁ Build the Dictionary using the rest of candidate part clusters.
  Optimize sparse representation coefficients:
  $\min_{A_j} \| X_j - D_{\bar{j}} \cdot A_j \|_2^2 + \lambda \| A_j \|_1$ ◁ Compute the reconstruction coefficients A_j
  $C_j \Longleftarrow Score\ (\| X_j - D_{\bar{j}} \cdot A_j \|_2^2)$ ◁ Compute the reconstruction residual for the j-th candidate part cluster
end for
Rank all the candidate part clusters using the obtained residual values.
Select the top K discriminative part clusters.
for $n = 1$ to K **do**
  $V_n \Longleftarrow LinearSVM\ (C_n)$ ◁ Learning the discriminative part detector
end for

Fig. 3 The top row shows some sample images. The bottom row demonstrates the corresponding probability maps.

patches. For all the patchs of image, the patch descriptors (HOG [11]) are extracted and we use the set of learned discriminative part detectors to score them. Then all the scores obtained by the part detectors are accumulated to generate the probability map. This probability map is further refined by the constraint of local regions. In our approach, we use the over-segmentation method [14] to generate these local superpixel regions and all the pixels have the same probability value within each superpixel (we use the average probability value in a local superpixel region). The higher probability value of a region, the more possibility that this region covers an object of interest. Figure 3 shows some instances and the corresponding probability maps. Notice that the obtained probability map can highlight the object of interest in an image and provide an important cue to detect the object.

Once we obtain the probability map, we use the method [7] to generate these candidate windows in an image, then each candidate window can be evaluated by the map. Specifically, the score for a window is measured by

$$F(W) = \frac{N_p^+(W)}{N_p^-(W)} - \frac{N_p^+(W_S)}{N_p^-(W_S)} \tag{2}$$

where $F(W)$ denotes the score for the current window W, $N_p^+(\cdot)$ is the number of pixels that the probability value is larger than p, $N_p^-(\cdot)$ counts the pixels that the probability value is less than p. The W_S denotes the surrounding region for the current window W, which is defined by the extended rectangular ring with respect to W in four directions. In the experiment, the value of p is computed by the average of probability map and the extended range of each window boundary is set to 30 pixel width. The higher score of a window indicates that this window is more likely to contain an object of interest. Then all the candidate windows are ranked by the window scores and the non-maxima suppression (NMS) is applied to generate the final object windows.

4. Experiments

We evaluate the proposed object detection approach on the PASCAL VOC 08 [15] standard benchmarks, and compare it to other recent competing methods. The approaches we compare against include the deformable part-based model (DPM) [1], Objectness [5] and Selective search [7]. The performances of different methods are reported using the standard Average Precision (AP) measure.

Since the objects have a large variety of categories, appearances, deformations and viewpoints in the challenging PASCAL VOC dataset, we partition the total object categories into several super-classes to verify the effectiveness of proposed method. The object categories within each super-class may share the potential consistent parts in terms of geometric properties. In particular, we divide all the 20 object categories into four super-classes (e.g. <cow, horse, sheep>; <bus, train, plane, car, boat, bicycle, motor>; <dog, cat, bird, person>; <table, sofa, tv, chair, bottle, plant>), then the semantic part detectors are learned from each super-class automatically. In the test, the learned semantic part detectors are used to capture the objects of corresponding categories in each super-class. We use the dataset splits from [15] for training and testing in the experiment.

For our discriminative semantic parts learning, each object region by the bounding box is regularly divided by the fixed 3×3 grid size. Then we extract the dense HOG features for all the pixels within an object part region. The visualizations of HOG features for object parts are shown in Fig. 1. Finally, the part descriptor is constructed by the max-pooling technology. In the experiment, the number of discriminative part clusters K is set by $\tau\times$ the number of candidate part clusters M, where $\tau \in (0, 1)$ is a constant that

	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
DPM [1]	.371	.590	.185	.133	.234	.337	.588	.161	**.248**	.586	.236	.161	.588	.519	**.633**	.182	.472	.297	**.407**	.419	.367
Objectness [5]	.375	.585	.202	.146	.227	.345	.586	**.176**	.231	.596	.237	.160	.592	**.528**	.622	.179	.499	.300	.390	.425	.370
Selective search [7]	.361	.581	.204	.145	.248	.338	.589	.168	.234	.625	.226	.155	.586	.518	.629	.190	.516	.295	.395	.424	.372
Ours	**.379**	**.592**	**.212**	**.155**	**.251**	**.346**	**.606**	.173	.246	**.638**	**.250**	**.162**	**.597**	.526	.625	**.204**	**.522**	**.301**	.395	**.431**	**.381**

Fig. 4 Average Precision (AP) for all the 20 classes in PASCAL VOC 2008 dataset. The highest performance among the proposed approach and some representative methods are displayed in bold.

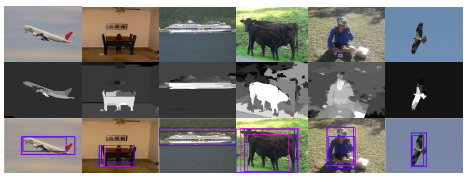

Fig. 5 First row: sample images in the PASCAL VOC 08 dataset. Middle row: probability map for each sample image. Third row: detection results by our method (blue) and the ground truth (magenta).

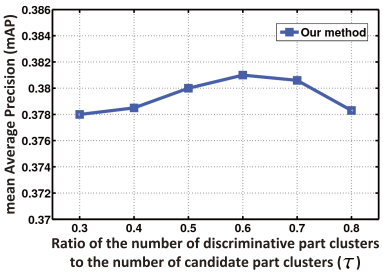

Fig. 6 The performance of proposed method with varying number of discriminative part clusters on the PASCAL VOC 08 dataset.

controls the number of discriminative part clusters. The parameter τ is set to 0.6 in the proposed method. In addition, the regularization parameter of Eq. (1) is set as $\lambda = 0.1$.

Note that the DPM [1] is one of the most successful class-specific object detection methods. It achieves the state-of-the-art performance and is hard to be further promoted. Similar to the recently proposed Objectness [5] and Selective search [7] algorithms, our approach is essentially a class-generic object detection method. Following the same settings in Objectness [5], we linearly combine the score $F(W)$ of our class-generic object detector with the DPM score $D(W)$ for each candidate window W: $D(W)+\alpha \cdot F(W)$, where the weighting parameter α is set to 0.2. Different from the Objectness [5], the method of [7] uses the trained classification model per class to score each candidate window for addressing the object recognition task. The detection results of different methods for all the 20 categories on PASCAL VOC 08 dataset are listed in Fig. 4. It shows that our method improves the AP and obtains the best performances for 15 out of 20 object classes, such as *plane, bicycle, bird, boat, bottle, bus, and etc*. Compared with the DPM, it is noticed that the proposed method increases the mean Average Precision (mAP) by more than one percent. Moreover, our approach outperforms other competing methods and gains the highest mAP across all the classes. Figure 5 illustrates some examples of visual results by the proposed method on VOC 08 dataset. The number of discriminative part clusters can influence the performance of our method. Therefore, we also evaluate the mAP of proposed method with different number of discriminative part clusters in Fig. 6.

5. Conclusion

In this letter, we propose a new discriminative parts learning approach to tackle the object detection problem with only the bounding boxes of object category labels. The proposed approach has the ability to discover the mid-level semantic part set and learn object part detectors automatically. Experimental results verify that the proposed approach can achieve superior performance to recent competing methods.

Acknowledgments

This work was supported in part by the program for Science and Technology Innovative Research Team for Young Scholars in Sichuan Province, China (No. 2014TD0006), and the Jiangsu Province Science Foundation for Youths, China (No. BK20130107).

References

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, 2010.

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

[3] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," Proc. CVPR, pp.2642–2649, 2013.

[4] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," Proc. ICCV, pp.1365–1372, 2009.

[5] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.11, pp.2189–2202, 2012.

[6] H. Li and K.N. Ngan, "Learning to extract focused objects from low dof images," IEEE Trans. Circuits Syst. Video Technol., vol.21, no.11, pp.1571–1580, 2011.

[7] K.E.A. Van de Sande, J.R.R. Uijlings, T. Gevers, and A.W.M. Smeulders, "Segmentation as selective search for object recognition," Proc. ICCV, pp.1879–1886, 2011.

[8] H. Li and K.N. Ngan, "A co-saliency model of image pairs," IEEE Trans. Image Process., vol.20, no.12, pp.3365–3375, 2011.

[9] H. Li, F. Meng, and K.N. Ngan, "Co-salient object detection from multiple images," IEEE Trans. Multimedia, vol.15, no.8, pp.1896–1909, 2013.

[10] B.J. Frey and D. Dueck, "Clustering by passing messages between data points," Science, vol.315, no.5814, pp.972–976, 2007.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. CVPR, pp.886–893, 2005.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation,"

IEEE Trans. Signal Process., vol.54, no.11, pp.4311–4322, 2006.
[13] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," Proc. ICML, pp.1–8, 2009.
[14] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graph-based image segmentation," Int. J. Comput. Vision, vol.59, no.2, pp.167–181, 2004.
[15] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org