Mining sub-categories for object detection

Jifeng Dai, Jianjiang Feng and Jie Zhou
State Key Laboratory on Intelligent Technology and Systems, TNList
Department of Automation, Tsinghua University, Beijing, China, 100084
djf05@mails.tsinghua.edu.cn, {jfeng,jzhou}@tsinghua.edu.cn

Abstract

The visual concept of an object category is usually composed of a set of sub-categories corresponding to different sub-classes, perspectives, spatial configurations and etc. Existing detector training algorithms usually require extensive supervisory information to achieve a satisfactory performance for sub-categorization. In this paper, we propose a detector training algorithm which can automatically mine meaningful sub-categories utilizing only the image contents within the training bounding boxes. The number of sub-categories can also be determined automatically. The mined sub-categories are of medium size and could be further labeled for a variety of applications like sub-category detection, meta-data transferring and etc. Promising detection results are obtained on the challenging PASCAL VOC dataset.

1. Introduction

The basic detectors are indispensable components to fill the sematic gap between the raw pixels and the symbolic token representations in complex structural object models [11, 9, 10, 6]. Different detectors are trained to capture the visual clues of different object categories, parts, surroundings and etc. The accuracy of the trained detector is strongly affected by the quality of sub-categorization. For example, if the training algorithm mixes dogs of different kinds, perspectives or poses together, then it is quite unlikely to produce a good detector of dog. Besides, a variety of applications could be explored beyond object detection once we know the sub-category of the detected object. The detector could answer questions like what kind of dog is in the picture, whether the dog is running and what the perspective is instead of a simple bounding box.

The key challenge lies in that we do not know the appropriate number of sub-categories in a category and which examples should be assigned to which sub-category. Due to the disturbance like complex background and varying illumination conditions, directly clustering based on the image content is of limited help. Some works attacked this problem by pre-fixing some parameters and introducing auxiliary information. For example, in [11, 1], images are extensively labeled so as to derive meaningful sub-categories by clustering the labeling information. In [5] and its derivations [10, 6], the number of sub-categories is pre-fixed at a small number (3 in [5]) before training, and the bounding box ratio information in the PASCAL VOC dataset [4] is utilized to get the initial assignment. Then a latent support vector machine (LSVM) is trained. However, the bounding box ratio information is not quite useful for some

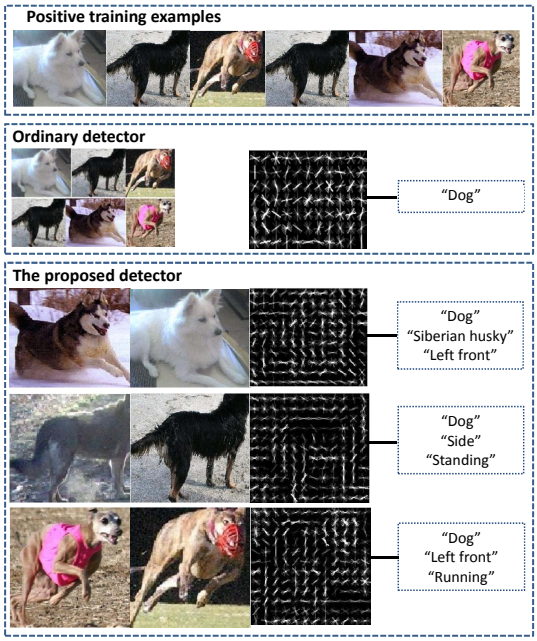

Figure 1: Ordinary detector vs. the proposed detector. The visual sub-categories mined for the proposed detector are much more meaningful and the number of sub-categories is determined automatically. The mined sub-categories could be further labeled to perform meta-data transferring.

Preprint submitted to 21st International Conference on Pattern Recognition.
Received April 7, 2012.

challenging categories, like bird. In [8, 2], sub-category number and training example assignment are both prefixed by assigning each positive example to a different sub-category. And the produced detectors could be prohibitively large on large training datasets (for example, there are 4690 positive training examples for the person category in the PASCAL VOC 2007 dataset).

In this work, we propose an object detector training algorithm that mines sub-categories efficiently. It optimizes sub-category number, training example assignment and detector parameters under a coherent framework without using any axillary information. An illustration of a dog detector trained on the challenging PASCAL VOC 2007 dataset is shown in Fig. 1, in which the mined sub-categories are very meaningful. The major contributions of this work include:

- The training problem is formulated as training a LSVM with adaptive number of components. The number is automatically determined by how many meaningful visual sub-categories could be mined for a category, making the algorithm have large flexibility to deal with various object categories.
- A convergent iterative influence expanding algorithm is proposed, optimizing the sub-category number, training example assignment and detector parameters under a coherent framework. The influence expanding training process could efficiently mine meaningful sub-categories in spite of the highly non-convex nature of the problem.
- The algorithm utilizes only the image content in the bounding box for training without requiring any auxiliary information. The problem of clustering training examples is solved in the training process. This makes the proposed approach broadly applicable for different settings.

For demonstration, we trained basic detectors on the challenging VOC 2007 dataset [4] without the aid of any auxiliary supervision information, structural model or contextual rescoring technology. Experiments show that the trained basic detectors can achieve a comparable accuracy with those of state of the arts [5, 8], in which complicated structures or contextual information is utilized.

2. Detecting algorithm

Given an input image, we adopt the standard sliding window procedure to detect instances of a category [3]. The detection window is scanned across the image at all locations and scales. For each location at each scale, a D-dimensional HOG [3] feature vector x is extracted and a score is computed by the detector of a certain category

$$f_W(x) = \max_{c=1,...,k} w_c^T x. \quad (1)$$

The detector W consists of k linear filters, $\{w_1, w_2, ..., w_k\}$ and each filter/component is a D-dimensional feature vector, which corresponds to a visual sub-category, such as cars of a certain type or a certain viewpoint.[1] The score of x is equal to the highest response of all k components.

The windows whose scores are above a threshold are detected as candidates, and non-max suppression is used to remove redundant detections. In the next section, the detector training algorithm is described.

3. Training Algorithm

Given a set of positive and negative training examples, $D^+ = \{x_1^+, \ldots, x_n^+\}$ and $D^- = \{x_1^-, \ldots, x_m^-\}$, the objective function is:

$$\begin{aligned} L(W) =& \|W\|_{2,1} + C\{ \sum_{i \in D^+} \max(0, 1 - f_W(x_i^+)) \\ &+ \sum_{j \in D^-} \max(0, 1 + f_W(x_j^-))\}, \end{aligned} \quad (2)$$

where $\max(0, 1 - f_W(x_i^+))$ and $\max(0, 1 + f_W(x_j^-))$ are the standard hinge losses and the constant C controls the tradeoff between model compactness and representability. $\|W\|_{2,1}$ is the 2,1-norm enforced on W, which takes the form $\|W\|_{2,1} = \frac{1}{2}\sum_{c=1}^{k} \|w_c\|_2$. This formulation is different from those in [5, 6, 10] in two aspects. First, the number k of components is not a pre-fixed constant, but is a variable to be optimized in training, since we do not know how many meaningful components a category contains. Second, mixed 2,1-norm is enforced to encourage group sparsity [7]. The form of the classification function in (1) naturally divides the variables in W into k non-overlapping groups $\{w_1, w_2, ..., w_k\}$. And we make use of this group structure so that the variables of the same component tend to be zeros or nonzeros simultaneously.

Despite these advantages, optimizing the function in (2) is very challenging. The training algorithm has to decide the assignment of the i-th positive example c_i, the number of components k and component parameters W. This makes the problem highly non-convex and hard to optimize. The traditional approach fixes component number k and utilizes an EM step to alternatively

[1] In this paper, the two terms "component" and "sub-category" are used interchangeably.

Preprint submitted to 21st International Conference on Pattern Recognition.
Received April 7, 2012.

Table 1: Procedure of the Influence Expanding Training Algorithm.

1 Initialize example assignments $A_c^{(0)} := \{c\}$, in which $c = 1, \ldots n$, and set $I^{(0)} = [1, \ldots, 1]$
2 Optimize initial components: $W^{(0)} := \arg\min_W L(W, I^{(0)}, A^{(0)})$
3 **for** $t := 0$ to T **do**
4 Generate new assignments $A^{(t+\frac{1}{2})}$ by influence expanding, and set $I^{(t+\frac{1}{2})} = [1, \ldots, 1]$
5 Optimize candidate components: $W^{(t+\frac{1}{2})} := \arg\min_W L(W, I^{(t+\frac{1}{2})}, A^{(t+\frac{1}{2})})$, and set $W^{(t+1)} := [W^{(t)} \vert W^{(t+\frac{1}{2})}]$
6 Component pruning: Derive $\{I^{(t+1)}, A^{(t+1)}\}$ by minimizing $L(W^{(t+1)}, I, A)$ greedily, subject to the constraints in (4)
7 Stop iteration if $L(W^{(t+1)}, I^{(t+1)}, A^{(t+1)})$ reaches a local optimum
8 **end**

optimize c_i and W [5, 10, 6]. Since it tries to minimize the objective function in each step, the approach can easily get stuck at local minimum. So initialization is carefully performed by fixing k at a small number and clustering the positive examples into k clusters using aspect ratio information.

We proposed an influence expanding training algorithm, optimizing assignment and number of components simultaneously. At each iteration, new candidate components are created by influence expanding, forming a redundant component set together with the existing components. Next, redundant and weak components are deleted by component pruning to directly minimize the objective function. In the influence expanding process, more components than necessary are created. And this gives a rise to the objective function at each iteration round, efficiently avoiding been trapped at some bad local optimal. To perform the above mentioned training, we define an auxiliary optimization problem:

$$L(W, I, A) = \sum_{c=1}^{\infty} L(w_c, A_c) \cdot I_c, \tag{3}$$

$$\text{s.t.} \quad \cup_{c=1}^{\infty} A_c \cdot I_c = D^+, \tag{4}$$

in which the maximum number of components is set to infinity to facilitate the adding of new candidate components, I_c is introduced to indicate whether the c-th component is selected ($I_c = 1$) or not ($I_c = 0$), A_c denotes the set of positive examples assigned to the c-th component. Note that the hinge loss for negative examples $\max(0, 1 + f_W(x_i^-))$ is convex in W, so we could simply assign negative examples to all the components. $L(w_c, A_c)$ takes the form:

$$\begin{aligned} L(w_c, A_c) = &\frac{1}{2}\|w_c\|_2 + C\{\sum_{i \in A_c} \max(0, 1 - w_c^T x_i^+) \\ &+ \sum_{j \in D^-} \max(0, 1 + w_c^T x_j^-)\}. \end{aligned} \tag{5}$$

The constraint (4) is added to make sure that $L(W, I, A)$ is an upper bound for the objective function. So we could train the LSVM with adaptive number of components by minimizing $L(W, I, A)$.

The procedure of the proposed training algorithm is shown in Table 1. Since there is no axillary information, we initialize from assigning each positive example to a different component. Next, the initial components are trained by optimizing $L(W, I^{(0)}, A^{(0)})$, in which $I^{(0)} = [1, \ldots, 1]$. Note that this is a convex function of W and could be optimized efficiently.

At each iteration, new assignments $A^{(t+\frac{1}{2})}$ are generated by influence expanding. Each selected component $w_c^{(t)}$ $(I_c^{(t)} = 1)$ expands its influence to the highest responded positive example out of its influence range and produces a new assignment $A_c^{(t+\frac{1}{2})}$. In contrast to the greedy positive example assigning schedules [5, 10, 6], our algorithm generates a set of alternative possible assignments, forcing the algorithm to explore more possible solutions. By adding a new positive example with the highest response, it is likely that the newly added one is similar to the existing ones, producing a high quality assignment. Then the corresponding candidate components $W^{(t+\frac{1}{2})}$ are trained and merged with the existing ones to form a new candidate component set $W^{(t+1)} = [W^{(t)} \vert W^{(t+\frac{1}{2})}]$.

Next, the redundant and weak components are pruned. This is performed by minimize $L(W^{(t+1)}, I, A)$ as a function of I and A. We optimize it greedily by first deleting the component which causes the largest loss, and then the one causes the second largest, till the objective function will deteriorate if we delete any more components. In the above process, A is set by assigning each example to the component giving the highest response.

Finally, the iterative training stops when it reaches a local optimal or the max iteration round is achieved.

4. Experiments

4.1 Experimental settings

We evaluate our basic detector training algorithm on the well-established PASCAL VOC 2007 dataset [4], which is widely acknowledged as a difficult testbed for object detection. Due to limited computing resources, we trained basic detectors for six animal categories, including bird, cat, cow, dog, horse and sheep.[2] The parameter C is set as 0.3. The images in the trainval set are utilized for training. The Average Precision (AP)

[2]Our code and trained models are available online at `http://ivg.au.tsinghua.edu.cn/`.

Preprint submitted to 21st International Conference on Pattern Recognition.
Received April 7, 2012.

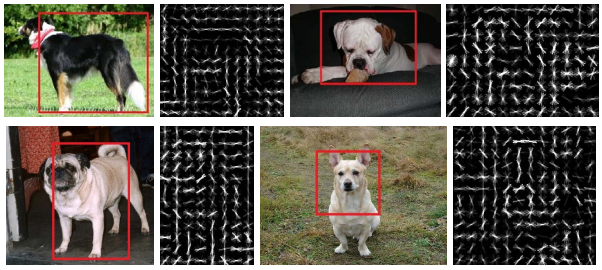

Figure 2: Some detection examples of the dog category on the PASCAL 2007 dataset.

and mean Average Precision (mAP) scores are calculated across the test set for evaluation, following the standard protocol in [4].

4.2 Comparison with Dalal/Triggs method

The algorithm most similar to the proposed one is the Dalal/Triggs (DT) [3] method, which trains a basic detector with a single sub-category for each category. Both algorithms utilized only basic detectors without any axillary supervisory information, structural models and contextual rescoring technology. The comparison of the proposed algorithm with the DT method is shown in Table 2. It can be seen that the proposed algorithm outperforms the DT method by a clear margin. The reason lies in that our trained detectors correspond to much more meaningful sub-categories. Some detection examples are shown in Fig. 2.

4.3 Comparison with state-of-the-arts

We also compared with state-of-the-art Latent Deformable Part Model (LDPM) [5] and the exemplar-SVM model (E-SVM) [8], as shown in Table 3. In [5], bounding box ratio prior, part-based models and contextual information are all utilized to enhance the performance. And we achieve mean Average Precision (mAP) on par with it using only simple basic detectors. In addition, our detectors could be further labeled to answer detailed questions like what kind of dog is in the picture, whether it is running and what the perspective is, while the detectors in [5] could not (see Fig. 1).

The mAP score of our detectors is also comparable with that in [8], in which contextual rescoring technology is utilized. An important advantage of our work is that the trained detectors are much more compact. The sub-category numbers of our detectors are $4 \sim 14$ times less than those in [8]. And this brings the benefit of much faster detectors and much less labor when further labeling the mined sub-categories.

Table 2: Average Precision (AP) scores and mean AP (mAP) of the proposed algorithm and the DT [3] method across the animal categories at the PASCAL VOC 2007 dataset.

Approach	bird	cat	cow	dog	horse	sheep	mAP
Proposed algorithm	**.094**	**.104**	**.140**	**.099**	**.420**	**.137**	**.166**
DT [3]	.005	.005	.128	.004	.122	.056	.053

Table 3: AP scores, mAP scores and sub-category numbers (in brackets) of the proposed algorithm and the state-of-the-art methods [5, 8] across the animal categories at the PASCAL VOC 2007 dataset.

Approach	bird	cat	cow	dog	horse	sheep	mAP
Proposed algorithm	**.094** (34)	.104 (76)	.140 (52)	**.099** (107)	.420 (86)	.137 (37)	.166
E-SVM [8]	.077 (486)	.052 (376)	**.186** (259)	.031 (510)	.447 (362)	**.226** (257)	**.170**
LDPM [5]	.006 (3)	**.163** (3)	.166 (3)	.050 (3)	**.452** (3)	.174 (3)	.169

5. Conclusions

We presented an algorithm of mining sub-categories to train basic detectors for object detection. The algorithm can mine meaningful sub-categories and produce detectors achieving accuracy on par with state-of-the-art results. We plan to study labeling the mined sub-categories for meta-data transferring.

References

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. *ICCV*, 2009.

[2] O. Chum and A. Zisserman. An exemplar model for learning object classes. *CVPR*, 2007.

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

[4] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 2010.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

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

[7] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 2010.

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

[9] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. *CVPR*, 2010.

[10] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. *CVPR*, 2010.

[11] S. Zhu and D. Mumford. A stochastic grammar of images. *FTCGV*, 2007.

Preprint submitted to 21st International Conference on Pattern Recognition.
Received April 7, 2012.