Discriminative Middle-Level Parts Mining for Object Detection

SUMMARY  Middle-level parts have attracted great attention in the computer vision community, acting as discriminative elements for objects. In this paper we propose an unsupervised approach to mine discriminative parts for object detection. This work features three aspects. First, we introduce an unsupervised, exemplar-based training process for part detection. We generate initial parts by selective search and then train part detectors by exemplar SVM. Second, a part selection model based on consistency and distinctiveness is constructed to select effective parts from the candidate pool. Third, we combine discriminative part mining with the deformable part model (DPM) for object detection. The proposed method is evaluated on the PASCAL VOC2007 and VOC2010 datasets. The experimental results demonstrate the effectiveness of our method for object detection.

key words: object detection, discriminative part mining, unsupervised part training, part selection

1. Introduction

Object detection is one of the key techniques in computer vision, which has broad applications in video surveillance, robotic vision, intelligent transportation, human-computer interaction, etc. However, it is still a challenging issue that aims at detecting generic objects that vary greatly in appearance. These variations arise from changes in viewpoint, subtype, occlusion and non-rigid deformation.

The deformable part model (DPM) \[1\] is one of the most popular object detection models. A deformable part model of a specific object category is represented by several components which represent subcategories, and each component is composed of a root model which represents the entire object and several part models in deformable configurations. DPM handles appearance variations and has achieved competitive performance on the PASCAL VOC benchmark datasets. Many efforts \[12\], \[13\], \[18\]–\[20\], \[22\], \[29\] have been made to improve DPM recently. However, the performance of DPM is still limited by inferior part search. In detail, DPM performs greedy part search for part model initialization under an energy maximization criterion, which may not be optimal for some objects. In an earlier work \[20\], part search was simplified using spatial pyramid structure. Although deeper hierarchies can capture finer information in the tree model, background regions or suboptimal parts may be selected. And-or tree (AOT) models \[22\] learn parts configurations by a dynamic programming algorithm. Non-rectangular part discovery \[29\] is proposed for DPM by exploiting object structures. In general, existing DPM based methods directly derive parts from the root model. Therefore, the discriminability of parts has not been well studied.

To address these limitations, we propose a discriminative part mining method for object detection in this paper. Our method mainly consists of three steps: initial parts generation, part detectors training and discriminative parts selection. First, following DPM, we learn the root model for each component and use selective search to generate initial parts from top-scoring objects by the root model. Second, we train part detectors by exemplar SVM. Third, we select effective parts from the candidate pool by a part selection model based on consistency and distinctiveness. Consistency is defined to represent the appearance similarity among the same parts while distinctiveness is defined to represent the classification ability of a part detector for distinguishing objects from background. We combine discriminative part mining with DPM for object detection and evaluate its performance on the PASCAL VOC2007 and VOC2010 datasets. The experimental results demonstrate the effectiveness of our method. An overview of the proposed discriminative part mining method is shown in Fig. 1.

To summarize, this work features three aspects:

• We introduce an unsupervised, exemplar-based training process for part detection.
• We have constructed a part selection model based on consistency and distinctiveness for part selection.
• We combine discriminative part mining with DPM for object detection and achieve superior performance on the challenging PASCAL VOC datasets.

The rest of this paper is organized as follows. In Sect. 2, we make a literature survey on DPM based methods and part discovery methods. In Sect. 3, we introduce our method on discriminative part mining, including initial parts generation, part detectors training and discriminative parts selection. In Sect. 4, we introduce our part models for object detection. The experimental results are shown in Sect. 5. Finally, conclusions are drawn in Sect. 6.

2. Related Work

Deformable part model based methods. Regarded as an influential model for object detection, DPM has received great attention and many efforts have been made to improve
Divvala et al. [10] demonstrate that considerable improvement can be obtained by more subcategories and better initialization in DPM. Using more powerful features or integrating more features can usually improve the performance. Zhang et al. [12] propose a boosted local structured HOG-LBP based object detector. Filder et al. [18] exploit semantic segmentation to help object detection. Ren et al. [13] replace HOG features with HSC features in DPM and achieve better performance. On the other hand, some methods attempt to explore better part configurations in DPM. Zhu et al. [20] construct a three-layer spatial pyramid structure. Although deeper hierarchies can capture finer information in the tree model, background regions or suboptimal parts may be selected. Zhou et al. [29] discover non-rectangular parts by exploiting object structures. Song et al. [22] propose an and-or tree model to learn part configurations. These existing methods derive parts from the root model, which may not be optimal. In contrast, we mine discriminative middle-level parts from training data and the discriminability of parts is improved in DPM.

Part discovery methods. The notion of part is important to computer vision because much recent work relies on the idea of representing an image/object as a composition of middle-level parts. Given additional annotations of parts or keypoints, some methods [19], [24] directly learn part models for object detection. However, part labeling requires painstaking manual work, and not all of the labeled semantic parts are discriminative for recognition. Therefore many works on unsupervised part discovery has been explored [4]–[9], [28], [30]. Some use clustering to mine parts. In [6], [7], a set of parts are mined by unsupervised discriminative clustering. Doersch et al. [30] develop an extension of mean-shift algorithm based on clustering to mine parts. Jain et al. [8] use an exemplar-based clustering to mine spatio-temporal parts. Clustering based approaches can discover consistent parts but they rely on standard distance metrics, which may not work well in high-dimensional feature space. Others use boosting to select parts. Viola et al. [28] select a small number of haar-like features from a large set by AdaBoost algorithm. Wang et al. [4] select a set of regions by a cascaded boosting classifier. Boosting based approaches can discover complementary parts through learning weak classifiers and adding them to strong classifiers iteratively. However, these approaches are affected by the complexity of boosting structure and training examples, which may cause overfitting or low generalization ability. In addition, some methods propose task-specific criteria for part evaluation and selection. Endres et al. [5] use forward selection to iteratively choose parts with the average max precision measure. Juneja et al. [9] propose entropy-curves as a means of evaluating and selecting parts. In contrast, we construct an part selection model based on consistency and distinctiveness for discriminative part mining.

3. Discriminative Part Mining

Unsupervised part discovery is a challenging problem because (1) it is hard to judge whether a part occurs in an image/object or not, and (2) it is hard to determine the location of the part when it occurs, and (3) it is hard to select effective parts located in appropriate layouts. We propose a discriminative part mining method for object detection in this paper. Our method mainly consists of three steps: initial parts generation, part detectors training and discriminative parts selection.

3.1 Initial Parts Generation

Without prior knowledge, parts are likely to appear in any sub-window of any training image. One could simply generate initial parts by exhaustive search. Unfortunately, it will lead to high computational complexity and most of these parts are not discriminative. Strategies based on interest points or regions are commonly used for sampling. However, a generic interest point detector such as Harris Corner or DoG operator defines a keypoint by means of only grayscale features. Moreover, the local patch centered on the keypoint is not appropriate for a middle-level part due to the limited size in DPM.

We use instead interest regions generated by selec-
ative search [15] to generate initial parts. Starting from an oversegmentation, selective search takes a bottom-up hierarchical grouping algorithm based on multiple low-level image features to form regions. Three strategies (single/fast/quality) are proposed in terms of different combinations of diversification configurations including a variety of color spaces, similarity measures and starting regions.

Given a set of training examples \(D = \{(x_i, y_i)\}_{i=1}^m\) where \(y_i \in \{-1, 1\}\) indicates the example \(x_i\) is an object or background, we first train the root model for each component by latent SVM. The component label \(c\) is treated as the latent variable and each example \(x_i\) is scored by:

\[
f_B(x_i) = \max_{1 \leq i \leq m} \omega_0^c \cdot \phi(x_i) + b_0^c
\]

(1)

where \(\phi(\cdot)\) denotes the HOG features, \(\beta_0^c = (\omega_0^c, b_0^c)\) denotes the root model parameter for component \(c\), and \(B = (\beta_0^1, \beta_0^2, \ldots, \beta_0^m)\) denotes the parameters for all components. The model parameters \(B\) are learned by minimizing the objective function:

\[
L_D(B) = \frac{1}{2} \|B\|^2 + C \sum_{i=1}^a h(f_B(x_i))
\]

(2)

where \(h(\cdot)\) denotes the hinge loss function and \(C\) controls the relative weight of the regularization term. Then, for each component, we choose \(M\) top-scoring examples as seeds and generate a set of initial parts \(R = \{r_i\}\) from these seeds by selective search.

### 3.2 Part Detectors Training

For each initial part, we use exemplar SVM to train a candidate part detector. Exemplar SVM [14] is an effective discriminative approach to detect similar examples with a specific example that is called an exemplar. In contrast with conventional SVM, exemplar SVM takes the exemplar as the only one positive example and uses different regularization parameters to control the relative weight between positive and negative examples.

In order to train part detectors, we take each initial part \(r_i\) as the positive example. For negative examples, we randomly sample windows from images that do not contain objects. These sampled windows form the negative training set \(N\). Each candidate part detector is trained by minimizing the following objective function:

\[
L(\beta_i) = \frac{1}{2} \|w_i\|^2 + C_1 h(w_i^T \cdot \phi(r_i) + b_i)
\]

(3)

\[
+ C_2 \sum_{r \in N} h(-w_i^T \cdot \phi(r) - b_i)
\]

where \(\beta_i = (w_i, b_i)\) denotes the parameter of the \(i\)th candidate part detector.

### 3.3 Discriminative Parts Selection

Not all of the candidate parts are discriminative. Therefore we define two measurements: consistency and distinctiveness, to select discriminative parts from the candidate pool.

For each component, we first detect parts on both positive examples \(P = \{I_i\}_{i=1}^{n_p}\) and negative examples \(N = \{I_i\}_{i=1}^{n_n}\). Then, for part \(i\), we define consistency as the mean value of the best detection scores on positive examples:

\[
\text{consistency}(i) = \frac{1}{n_p} \sum_{j=1}^{n_p} \text{part-detect-best}(\beta_i, I_j)
\]

(4)

and define distinctiveness as the log value of the ratio of the number of detections on positive examples to the number of detections on negative examples:

\[
\text{distinctiveness}(i) = \ln \frac{\sum_{j=1}^{n_p} \text{part-detect-all}(\beta_i, I_j, t)}{\sum_{j=1}^{n_n} \text{part-detect-all}(\beta_j, I_j, t)}
\]

(5)

The function part-detect-best(\(\beta_i, I\)) finds the highest scoring part hypothesis on the example \(I\) with the candidate part detector \(\beta_i\) and returns the confident score. The function part-detect-all(\(\beta_i, I, t\)) finds the best part hypotheses on the example \(I\) with the candidate part detector \(\beta_i\) above threshold \(t\) and returns the number of part detections. According to the above two definitions, consistency is computed to represent the appearance similarity among the same parts, and distinctiveness is computed to represent the classification ability of a part detector for distinguishing objects from background. Figure 2 shows the best and worst bicycle parts in terms of the two measurements, respectively.

Finally, a part selection model is constructed using a linear combination of consistency and distinctiveness after scaling to the equal range:

\[
F(i) = A_1 \cdot \text{consistency}(i) + A_2 \cdot \text{distinctiveness}(i)
\]

(6)

All candidate parts are ranked in terms of the value of \(F(i)\) and we select a set of top-ranked ones as the discriminative parts for object detection.

### 4. Part Models for Object Detection

**Part initialization.** In contrast with greedy search in DPM, we use spatial voting to initialize part models by the mined discriminative parts. For each part model, we create a two-dimensional array \(A\) to record the locations of the detected parts on objects. Initially, all the elements in the array \(A\) are
set to zero. Then we accumulate $A$ by

$$A(x, y) = A(x, y) + s$$  \hspace{1cm} (7)

If we take $s$ as the fixed value 1 in Eq. (7), it becomes hard spatial voting (equivalent to spatial locations clustering). If we take $s$ as the score of part detection, it becomes soft spatial voting. The optimal position (anchor) of the initial part model relative to the root model is calculated by maximizing $A$

$$(x^*, y^*) = \arg \max_{x,y} A(x, y)$$  \hspace{1cm} (8)

Figure 3 illustrates the part initialization procedure by spatial voting.

**Part Updating.** Following DPM, we train part models by latent SVM. The component label, the root location and the part locations are treated as latent variables in the training process. However, the anchor positions of part models are fixed throughout training process in DPM. Such part models are not flexible enough for object detection. In contrast, we apply spatial mixture modeling to update the anchor positions of part models during training. At each iteration, we first detect parts by latent SVM. Then we cluster the locations of the detected parts on objects and take the clustering centers as the new anchor positions.

**5. Experiments and Evaluations**

Following recent works on generic object detection, we evaluate the performance of the proposed method on the challenging PASCAL VOC2007 and VOC2010 datasets [26]. The PASCAL VOC datasets contain 20 object categories including rigid objects such as car and deformable objects like person. The available annotations contain the object category and the bounding box of each object without part information. We adopt the PASCAL VOC setup for evaluating detection performance. If a bounding box prediction overlaps the ground truth greater than $\tau = 0.5$ which is defined by the intersection over union (IoU), the detection is considered correct, otherwise it is false. Detection performance of each object category is measured by the average precision. The overall performance is reported by the mean of average precisions (mAP) over all object categories.

### 5.1 The Detection Results on the PASCAL VOC Datasets

We combine discriminative part mining with DPM for object detection. The models are trained with 6 roots and 8 parts per root. To validate the advantages of the proposed method, we compare it with three related methods: DPM-release 5 [27], and-or tree (AOT) models [22] and non-rectangular part discovery (NPD) [29]. DPM performs greedy part search for part model initialization. AOT models learn part configurations by a dynamic programming algorithm. NPD discovers non-rectangular parts by exploiting object structures.

Table 1 shows our detection performance compared with other methods on PASCAL VOC2007 dataset. Our method outperforms all the three methods in terms of mAP over all the 20 object categories. The experimental results demonstrate the effectiveness of discriminative middle-level parts.
Table 2  Performance comparison on the PASCAL VOC2010 dataset (average precision %).

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM [27]</td>
<td>42.9</td>
<td>47.2</td>
<td>10.3</td>
<td>11.1</td>
<td>26.3</td>
<td>48.4</td>
<td>40.2</td>
<td>22.9</td>
<td>17.0</td>
<td>22.9</td>
<td>10.2</td>
<td>19.9</td>
<td>41.5</td>
<td>44.0</td>
<td>41.0</td>
<td>7.6</td>
<td>28.3</td>
<td>18.2</td>
<td>39.0</td>
<td>32.9</td>
<td>28.6</td>
</tr>
<tr>
<td>AOT [22]</td>
<td>44.6</td>
<td>48.5</td>
<td>10.8</td>
<td>12.9</td>
<td>26.3</td>
<td>47.5</td>
<td>41.6</td>
<td>21.6</td>
<td>17.3</td>
<td>23.6</td>
<td>11.5</td>
<td>22.9</td>
<td>40.9</td>
<td>45.3</td>
<td>37.9</td>
<td>9.6</td>
<td>30.4</td>
<td>25.3</td>
<td>39.0</td>
<td>31.2</td>
<td>29.4</td>
</tr>
<tr>
<td>NPD [29]</td>
<td>45.9</td>
<td>50.7</td>
<td>10.4</td>
<td>11.2</td>
<td>28.7</td>
<td>50.5</td>
<td>44.2</td>
<td>24.2</td>
<td>17.4</td>
<td>24.0</td>
<td>13.7</td>
<td>17.6</td>
<td>40.2</td>
<td>45.7</td>
<td>38.8</td>
<td>8.3</td>
<td>29.3</td>
<td>20.2</td>
<td>41.0</td>
<td>35.6</td>
<td>30.0</td>
</tr>
<tr>
<td>Ours</td>
<td>47.4</td>
<td>52.2</td>
<td>10.5</td>
<td>14.5</td>
<td>32.9</td>
<td>50.2</td>
<td>47.8</td>
<td>23.7</td>
<td>15.0</td>
<td>24.9</td>
<td>10.4</td>
<td>9.6</td>
<td>43.0</td>
<td>45.5</td>
<td>42.6</td>
<td>7.9</td>
<td>29.0</td>
<td>19.8</td>
<td>40.4</td>
<td>35.6</td>
<td>30.1</td>
</tr>
</tbody>
</table>

Fig. 4  The validation results on bicycle (left) and cat (right) category with different values of $\lambda$.

Fig. 5  Some examples of the mined discriminative parts on the PASCAL VOC 2007 dataset.

Fig. 6  An example on the three strategies of selective search: the single strategy (top row), the fast strategy (middle row), and the quality strategy (bottom row). Five regions which have the highest overlap values with a semantic part (bounded yellow) are shown.

Table 3  The overlap values and the total number of regions in the example as shown in Fig. 6. The fast strategy generates high quality regions with modest amount.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Overlap values</th>
<th>#Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>single</td>
<td>0.26 0.23 0.20 0.00 0.00</td>
<td>47</td>
</tr>
<tr>
<td>fast</td>
<td>0.75 0.67 0.56 0.55 0.53</td>
<td>356</td>
</tr>
<tr>
<td>quality</td>
<td>0.89 0.86 0.77 0.76 0.75</td>
<td>1053</td>
</tr>
</tbody>
</table>

Fig. 7  The mined discriminative parts on a bicycle example with the different number of seeds. Some meaningful parts such as pedal and seat are missing in the case of single seed (M=1).

5.2 Parameter Analysis

We select discriminative parts by a part selection model based on consistency and distinctiveness. The actual values of consistency and distinctiveness are scaled to [0,1] by min-max normalization, respectively. Then the two measurements are combined in a linear manner. We study the effect of scale factor on detection performance in this experiment.

Scale factor $\lambda = \lambda_1/\lambda_2$. We test different values of $\lambda$ for combining consistency and distinctiveness, and apply the mined parts to DPM for object detection. The models are trained on the train set and tested on the validation set in this experiment. The best parameter of scale factor for each object category is selected by holdout validation. Figure 4 shows two validation results with different values of $\lambda$. The experimental results indicate that distinctiveness is more important for rigid objects and consistency is more important for deformable objects. Figure 5 shows some examples of the mined discriminative parts with the optimal $\lambda$ on the PASCAL VOC 2007 dataset.

5.3 Strategy Analysis

Selective search strategies. We test the three strategies of selective search (single/fast/quality) for generating initial parts. An example is shown in Fig. 6 and Table 3. The single strategy generates the least locations and the excessive sparsity may lead to dropping some meaningful regions. In contrast, the quality strategy generates the most locations but the excessive redundancy may increase the unnecessary computation time and burden the subsequent process of part
training and selection. The fast strategy generates high-quality regions with modest amount. Therefore we choose the fast strategy as the trade-off between the number and the quality of initial parts.

**Single seed vs. multiple seeds.** We test the different number of seeds for generating initial parts. The quality of initial parts has an impact on the performance of candidate part detectors since each initial part is taken as the training exemplar. If we only generate part proposals from the best example, it may lead to inferior parts because objects vary

---

**Fig. 8** The initial parts (left) and their HOG features (middle) and top 10 similar parts (right) using nearest neighbor search (top row for each part) and exemplar SVM (bottom row for each part).

**Fig. 9** The part detection results on a cat example by global search (top row) and local search (bottom row). The exemplar (left) and part detections are bounded red solid bounding box. The search scope of local search is bounded green dash bounding box.

**Fig. 10** Some detection examples of DPM (left) vs. our method (right), showing top 3 candidates (in the order of red, green, blue). From top to bottom, these object categories are bird, cat, chair, plant, train and tv, respectively.
greatly in appearance and not all of the parts generated from the single object example are representative. Furthermore, some effective parts will be missing when part occlusion or truncation occurs. Instead, we train a better part detector by choosing the best training exemplar from multiple seeds. Figure 7 shows the mined discriminative parts on an object example with the different number of seeds.

Nearest neighbor search vs. exemplar SVM. We test nearest neighbor search and exemplar SVM to find similar parts. For nearest neighbor search, cosine distance is used as the metric to judge correlation between parts. For exemplar SVM, hinge loss function is used and the regularization parameters are set as $C_1 = 0.5$, $C_2 = 0.01$. Figure 8 shows some examples of comparison results. For the wing of airplane, both approaches obtain good results because variations in appearance are not great for rigid objects. For the head of bird and the head and shoulder of person, some similar parts rank higher by exemplar SVM than nearest neighbor search. That indicates exemplar SVM is better for deformable objects. For the pot of pottedplant, 9 correct parts are found by exemplar SVM and only 5 correct parts are found by nearest neighbor search. That indicates although changes in subtypes and appearances result in great intra-class variations for indoor objects, exemplar SVM can obtain good part detections.

Global search vs. local search. We investigate two search schemes (global/local) to detect parts by exemplar SVM. By global search, we detect the best part in the whole object. By local search, we detect the best part in the surrounding area of the exemplar. Global search is a similarity-dominated scheme and it can capture better appearance similarity. While local search is a location-dominated scheme and it can capture better location aggregation. We choose global search to find parts with better appearance similarity and use spatial voting (discussed in Sect. 4) to model location aggregation. Figure 9 shows the part detection results on an object example by the two search schemes.

5.4 Run-Time Analysis

We conducted the experiments on a 3.5 GHz 4-core desktop computer running Ubuntu 14.04. Multiple threading is used to speed up the training and testing procedure. It takes about 8 hours to train an object model on the PASCAL VOC2007 trainval data set and 3 hours to evaluate it on the test data set. The total training time is divided into 4 hours for discriminative part mining and 4 hours for training DPM by latent SVM. In detail, it takes about 1 second to generate initial parts from an object example by the fast strategy of selective search. Training a candidate part detector by exemplar SVM takes about 1 second and detecting parts takes about 1~2 minutes. The consistency and distinctiveness are computed on the fly based on the scores of part detections. For testing, the average running time per image is about 2 seconds.

6. Conclusions

We have proposed an unsupervised method to mine discriminative middle-level parts from challenging datasets for object detection. We generate initial parts by selective search on object examples. Exemplar SVM is leveraged to train candidate part detectors to find similar parts. A part selection model is constructed based on consistency and distinctiveness for part selection. Consistency is defined to represent the appearance similarity among the same parts, capturing the idea that the more similar parts are, the higher the confident scores of part detections are. Distinctiveness is defined to represent the classification ability of a candidate part detector for distinguishing objects from background, capturing the idea that distinctive parts occur frequently in the positive examples and rarely in the negative examples. We combine discriminative part mining and DPM for object detection on the challenging PASCAL VOC datasets. Experimental results demonstrate the effectiveness of our method.

In the future work, we plan to develop an iterative learning algorithm that alternates between component learning and discriminative part mining. For some object categories which vary greatly in appearance such as plant, and object categories under severe occlusion such as table, our method may not work well due to the limitations of HOG features. Therefore we plan to learn more powerful CNN features for part mining. Also, applying discriminative part mining to scene classification and fine-grained category recognition is a natural extension.

Acknowledgments

This work was supported by the National High Technology Research and Development Program of China (863 program) under Grant No. 2012AA011004 and the National Science and Technology Support Program under Grant No. 2013BAK02B04.

References

References


Dong Li received the B.E. degree from Tsinghua University, China, in 2012. He is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include pattern recognition, computer vision, video analysis, etc.

Yali Li received the B.E. degree with Excellent Graduates Award from Nanjing University, China, in 2007 and the Ph.D degree from Tsinghua University, Beijing, China, 2013. Currently she is a research assistant in Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, and video analysis, etc.

Shengjin Wang received the B.E. degree from Tsinghua University, China, in 1985 and the Ph.D degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1997. From May 1997 to August 2003, he was a member of the research staff in the Internet System Research Laboratories, NEC Corporation, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently also a Director of the Research Institute of Image and Graphics (RIIG). He has published more than 50 papers on image processing. He is the holder of ten patents. His current research interests include image processing, computer vision, video surveillance, and pattern recognition. He is a member of IEEE and IEICE.