Non-rectangular Part Discovery for Object Detection

Chunluan Zhou
czhou002@e.ntu.edu.sg
Junsong Yuan
jsyuan@ntu.edu.sg

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Abstract

The deformable part-based model (DPM) is one of the most influential models for generic object detection and many efforts have been made to improve the model. Despite previous work, the problem of how to identify discriminative parts for DPM still remains largely unexplored. Most DPM based methods rely on a fixed number of parts of rectangular shapes, which may not be optimal for some object categories. In this paper, we present a novel approach to discover parts which can be non-rectangular by exploiting object structures. Instead of performing greedy part search as in DPM, our part discovery approach is carried out by first solving a K-way normalized cuts problem and then applying local refinement. Generally, the parts obtained by the proposed approach can better fit the object structures. We demonstrate the effectiveness of our approach on PASCAL VOC2007 and VOC2010 datasets.

1 Introduction

Although object detection has achieved great success on some specific object categories, *e.g.* human face [33], it is still difficult for existing methods to detect generic object categories that have a wide range of appearance variations. The pictorial structure [15] provides an expressive way to represent objects and has been used in some computer vision applications, such as object recognition [11], pose estimation [35] and action recognition [34]. One of the most successful applications of the pictorial structure is the deformable part-based model (DPM) [13] for object detection. A deformable part-based model of a specific object category is comprised by several components which represent different sub-categories and each component consists of a root which represents the entire object and several parts which can move relatively to the root to capture structural deformations. DPM can handle viewpoint changes and appearance variations to a large extent, and achieves competitive performance on challenging benchmark datasets, *e.g.* PASCAL VOC2010 [10].

Regarded as a promising model for generic object detection, DPM has received considerable attention from the object detection community and many efforts have been made to improve it, for example [1, 2, 4, 8, 16, 18, 19, 22, 31]. However, much less work has been done to discover parts for DPM. Most DPM based methods adopt the greedy search approach proposed in [13] to initialize a predefined number of parts of rectangular shapes, which may not be optimal for some object categories. Moreover, object structures are not well exploited

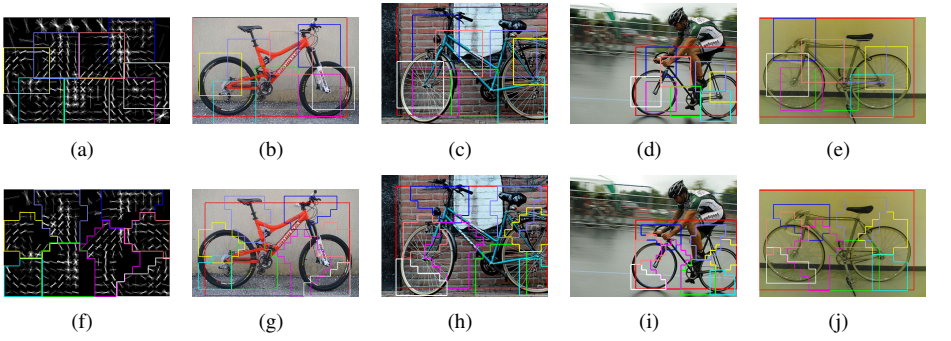


Figure 1: Rectangular parts vs. non-rectangular parts. (a) Part configuration discovered by the greedy search approach. (b-e) Detection examples of the part configuration (a). Red bounding boxes denotes the matching regions of the root filter. (f) Part configuration discovered by our approach. (g-j) Detection examples of the part configuration (f). It can be seen that the actual shapes of bicycle seat and handle are better fitted by non-rectangular parts and the part configuration (f) better matches the structure of a left/right facing bicycle.

by the approach. In [4, 38], a three-layer spatial pyramid structure is used to simplify the initialization of parts. An And-Or tree model [31] is proposed to select discriminative part configurations by a dynamic programming algorithm. Although the method can determine the part sizes for each component automatically, part shapes are still restricted to rectangles.

To address the limitations of the above part discovery approaches, we propose a novel data-driven approach to discover non-rectangular parts by exploiting object structures. Following DPM, we first learn the root filter for each component of the model and then derive parts from the root filter: the root filter is enlarged to twice its original size and a specified number of connected non-rectangular regions are cropped out of the enlarged root filter as parts such that the obtained part configuration well matches the structure of object examples. We implement this part discovery approach by solving a K-way normalized cuts problem [36] followed by local refinement. Compared to the greedy search approach of DPM, our approach has two advantages (See Fig. 1 for illustration):

1. The shapes of the parts discovered by our approach can be non-rectangular, which makes the obtained parts more suitable for matching non-rectangular regions;

2. Generally, part configurations obtained by our approach can better fit object structures than those obtained by the greedy search approach.

To evaluate our approach, we conduct object detection on two benchmark datasets, PASCAL VOC2007 [9] and VOC2010 [10]. The experimental results demonstrate the effectiveness of the proposed approach.

2 Related work

Generative graphic models, *e.g.* conditional random fields [28, 29] and k-fan statistical models [5, 24], are commonly used to model object structures. However, for object detection, these models often cannot compete with discriminatively trained models, *e.g.* DPM, on

benchmark datasets. Different from these structure modelling methods, our approach discovers parts in a discriminative way and provides initial parts for DPM to learn a better detection model. In recent years, many efforts have been made to improve DPM. Several methods combine different features, such as color attributes [19], scale-invariant feature transform (SIFT) [4], local binary pattern (LBP) [37] and segmentation features [14], with histogram of oriented gradients (HOG) [6] to improve the discriminative power of DPM. In [27], HOG is replaced by histograms of sparse codes (HSC) in DPM, achieving a significant performance gain. The detection process of DPM for a single object category is accelerated by employing KD-Ferns [22], leveraging transform [2], and adopting the cascade [12] or coarse-to-fine [26] detection framework. The recent work [7] exploits locality-sensitive hashing to efficiently detect a large number of object categories on a single machine. Building shared intermediate representations for part filters [20, 30] is another way to accelerate DPM for multiple object categories. Some sub-category clustering methods [1, 8, 18, 21, 31] are proposed to obtain better sub-categories for DPM. In [1, 16], flexible variants of DPM are introduced to handle partial occlusions. As shown in [3, 13, 32], contextual information can be made use of to improve the detection performance of DPM. In addition, category-specific information, like texture [25] and irregular patches [23], are exploited for refining the results obtained by DPM.

3 Non-rectangular Part Discovery

3.1 Deformable Part-based Model

The deformable part-based model (DPM) [13] of an object category consists of several components representing sub-categories of different orientations or poses. In the model, a component is represented by a root filter and some part filters, used for matching the whole region of an object and capturing finer details of the object, respectively. Part filters are allowed to move relatively to the root filter to handle structural deformations. Let M_c be the c-th component of the model with N_c part filters. The component M_c is defined by a $(2N_c+2)$-tuple $\beta_c = (\boldsymbol{F}_0, \boldsymbol{F}_1, ..., \boldsymbol{F}_{N_c}, \boldsymbol{d}_1, ..., \boldsymbol{d}_{N_c}, b)$, where $\boldsymbol{F}_0$ is the root filter, $\boldsymbol{d}_i \in \mathbb{R}^4$ is the deformation parameters of the part filter $\boldsymbol{F}_i$, and b is the bias term. Each filter $\boldsymbol{F}_i$ is an $H_i \times W_i$ array of n-dimensional weight vectors, where H_i and W_i are the height and width of $\boldsymbol{F}_i$, respectively. Denote by $P_c(X) = \{\boldsymbol{p_0}, \boldsymbol{p_1}, ..., \boldsymbol{p}_{N_c}\}$ a matching configuration of the the component M_c on an object example X, where $\boldsymbol{p}_i = (x_i, y_i)$ is the matching location of $\boldsymbol{F}_i$ on X. The matching score of the configuration $P_c(X)$ is defined by

$$S(\beta_c, P_c(X)) = \sum_{i=0}^{N_c} \boldsymbol{F}_i \cdot \phi(X, \boldsymbol{p_i}) - \sum_{i=1}^{N_c} \boldsymbol{d}_i \cdot \psi(dx_i, dy_i) + b, \tag{1}$$

where $\phi(X, \boldsymbol{p_i})$ denotes the HOG feature [6] extracted from the matching region of $\boldsymbol{F}_i$ at the location $\boldsymbol{p}_i$ on X and $\psi(dx_i, dy_i) = (dx_i^2, dx_i, dy_i^2, dy_i)$ denotes the deformation feature with (dx_i, dy_i) the displacement of $\boldsymbol{F}_i$ relative to its anchor position on the root filter. Assume the model has M components and denote the parameters of these components by $B = (\beta_1, ..., \beta_M)$. The matching score of the object example X for the model is defined by

$$H(B, X) = \max_{1 \le c \le M} \max_{P_c(X) \in Z_c(X)} S(\beta_c, P_c(X)), \tag{2}$$

where $Z_c(X)$ is a set of possible matching configurations of the component M_c on the object example X.

Denote by $D = \{(X_i, l_i) | 1 \leq i \leq N\}$ a set of training examples, where X_i is the image region of the i-th example and $l_i \in \{1, -1\}$ indicates whether the example X_i belongs to the specific object category. The model parameters B are learned by minimizing the following objective function

$$L(B,D) = \frac{1}{2} \| B \|^2 + C \sum_{i=1}^{N} \max(0, 1 - l_i H(B, X_i)). \tag{3}$$

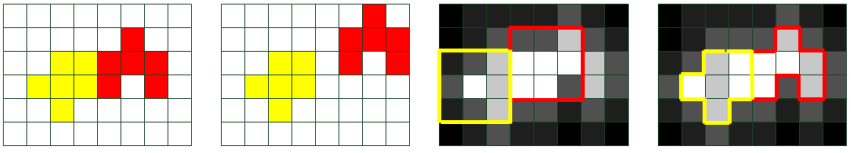
As sub-category labels and part annotations of training examples are not available, the model cannot be trained directly. Instead, training is done by first initializing the model parameters and then updating these parameters using the sub-category labels and part locations obtained by the initial model. In the initialization step, positive examples are first clustered into several sub-categories according to the aspect ratios of these examples. For each sub-category, a root filter is obtained by training a linear SVM classifier on positive examples belonging to the sub-category and negative examples randomly sampled from training images. Then, part filters are derived from the root filter. We will discuss how to initialize part filters in next section and refer readers to [13] for more details on model training.

3.2 Part Filter Initialization

The objective function in Eq. (3) is non-convex and, as reported in [13], the training process of DPM is susceptible to local minima, so it is necessary to select a good initialization of model parameters. We focus on the initialization of part filters which plays an important role in training the model. In [13], part filters of a component are initialized from its root filter in a greedy way: the root filter is enlarged to twice its original size by interpolation and several regions of predefined rectangular shapes (*e.g.* 6×6 squares) that have highest energy are cropped out from the enlarged root filter. The energy of a region is defined by the norm of positive weights corresponding to the region. Once a region is cropped out as a part filter, the weights corresponding to the region are set to zero and next highest-energy region is chosen until a specified number of part filters are obtained. This part filter initialization method has two limitations as illustrated in Fig. 2: first, the pre-defined rectangular shapes of part filters may not be optimal for the specific object category; second, the obtained part filters may not well fit the object structure. To obtain a better initialization of part filters, we propose a data-driven approach to discover part filters which can be non-rectangular by exploiting the structure of object examples.

Formulation. Let D_c be the set of object examples belonging to the c-th sub-category and $K = |D_c|$ be the number of object examples in D_c. For each component M_c, we aim to find N_c part filters that have good matching regions on object examples in D_c and are consistent with these examples in terms of object structure. First, we double the size of the root filter $\boldsymbol{F}_0$ by interpolation, as in [13], to capture finer details. The enlarged root filter, denoted by $\boldsymbol{F}'_0$, is represented by a $2H_0 \times 2W_0$ array of cells C_k for $1 \leq k \leq 2H_0 \times 2W_0$, where each cell C_k corresponds to a n-dimensional weight vector in $\boldsymbol{F}'_0$. Then, from $\boldsymbol{F}'_0$, we obtain a configuration of N_c connected part filters, $\Lambda = \{\boldsymbol{F}_i | 1 \leq i \leq N_c\}$, which satisfies the following overlapping constraint:

$$O(\boldsymbol{F}_i, \boldsymbol{F}_j) = \frac{Area(\boldsymbol{F}_i \cap \boldsymbol{F}_j)}{Area(\boldsymbol{F}_i \cup \boldsymbol{F}_j)} < \tau \quad \text{for } i \neq j, \tag{4}$$



(a) (b) (c) (d)

Figure 2: Toy example of part filter initialization. (a) An object example consisting of two parts. (b) A deformed object example whose two parts move away from each other. (c) The enlarged root filter with two 3×3 highest-energy part filters obtained by the greedy method. Bright cells indicate high-energy regions and dark cells indicate low-energy regions. (d) A better part-filter configuration which agrees with the structure of the two object examples.

where τ is an overlapping threshold. This constraint prevents any two part filters from overlapping largely. We measure the fitness of the part filter configuration Λ to object examples in D_c by

$$F(\Lambda) = S_R(\Lambda)^{\lambda} \times S_C(\Lambda), \tag{5}$$

where $S_R(\Lambda)$ is the average matching response of Λ over object examples in D_c, $S_C(\Lambda)$ reflects the structural consistency of Λ with these examples, and λ is a parameter used to balance $S_R(\Lambda)$ and $S_C(\Lambda)$. Our goal is to find a feasible part-filter configuration Λ that maximizes $F(\Lambda)$.

Now we describe how to evaluate $S_R(\Lambda)$. We first compute the matching response of each part filter $\boldsymbol{F}_i$ on a single object example X_j^+, $R_F(\boldsymbol{F}_i, X_j^+)$. The root filter $\boldsymbol{F}_0$ is applied to obtain its optimal matching region R_j on X_j^+ and the region R_j is enlarged to twice its original size. We extract HOG features, denoted by $\boldsymbol{f}_j$, from the enlarged region R'_j. Due to structural deformation and variation, R'_j may not be able to cover the whole region of the object example. In practice, we extend R'_j outward by a band of 2 cells width, so the final size of $\boldsymbol{f}_j$ is $(2H_0+4) \times (2W_0+4)$. The matching response $R_F(\boldsymbol{F}_i, X_j^+)$ can be computed by

$$R_F(\boldsymbol{F}_i, X_j^+) = \boldsymbol{F}_i \cdot \boldsymbol{f}_j(\boldsymbol{a}_i + \boldsymbol{\Delta}_i, \boldsymbol{F}_i), \tag{6}$$

where $\boldsymbol{a}_i$ is the anchor position of $\boldsymbol{F}_i$ (i.e. the top-left coordinate of $\boldsymbol{F}_i$ in $\boldsymbol{F}'_0$), $\boldsymbol{\Delta}_i$ is the displacement of $\boldsymbol{F}_i$ to its optimal matching region in $\boldsymbol{f}_j$ and $\boldsymbol{f}_j(\boldsymbol{a}_i + \boldsymbol{\Delta}_i, \boldsymbol{F}_i)$ denotes the HOG features extracted from the matching region and has the same shape as $\boldsymbol{F}_i$. We search for the optimal matching region of $\boldsymbol{F}_i$ in a neighborhood centered at $\boldsymbol{a}_i$ in $\boldsymbol{f}_j$ and do not take deformation penalty into account. With $R_F(\boldsymbol{F}_i, X_j^+)$, we define

$$S_R(\Lambda) = \frac{1}{K} \sum_{X_j^+ \in D_c} \sum_{\boldsymbol{F}_i \in \Lambda} R_F(\boldsymbol{F}_i, X_j^+). \tag{7}$$

Therefore, if part filters in Λ have good matching regions on object examples in D_c, the average response $S_R(\Lambda)$ should be high.

To evaluate $S_C(\Lambda)$, we consider the deformations between cells in $\boldsymbol{F}'_0$ on object examples in D_c. Let $\boldsymbol{\delta}_k^j$ be the displacement of the cell C_k to its optimal matching region on X_j^+. The optimal matching region is obtained in a neighborhood centered at (x_k+2, y_k+2) in $\boldsymbol{f}_j$, where (x_k, y_k) is the coordinate of C_k in $\boldsymbol{F}'_0$. To make the estimation of $\boldsymbol{\delta}_k^j$ more robust, we

search for the optimal matching region of the 3×3 block centered at C_k and take the corresponding displacement as $\boldsymbol{\delta}_k^j$. The deformation between cells C_k and C_l on X_j^+ is measured by $d_{kl}^j = \| \boldsymbol{\delta}_k^j - \boldsymbol{\delta}_l^j \|$. Letting $\bar{d}_{kl} = \frac{\sum_{j=1}^K d_{kl}^j}{K}$ be the average deformation between C_k and C_l over all object examples in D_c, we define the cohesiveness between cells C_k and C_l by

$$T(C_k, C_l) = \begin{cases} exp(-\frac{\bar{d}_{kl}^2}{\sigma^2}) & \text{if } C_k \text{ is adjacent to } C_l; \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where σ is a parameter and only the cohesiveness between adjacent cells is considered. The total cohesiveness between cells within $\boldsymbol{F}_i$ and the total cohesiveness between cells in $\boldsymbol{F}_i$ and cells in $\bar{\boldsymbol{F}}_i = \boldsymbol{F}_0' \setminus \boldsymbol{F}_i$ are given by

$$T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) = \sum_{C_k, C_l \in \boldsymbol{F}_i} T(C_k, C_l), \tag{9}$$

and

$$T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i) = \sum_{C_k \in \boldsymbol{F}_i, C_l \in \bar{\boldsymbol{F}}_i} T(C_k, C_l), \tag{10}$$

respectively. A cohesive part filter should have strong cohesiveness between its cells (i.e. large $T_W(\boldsymbol{F}_i, \boldsymbol{F}_i)$) and relatively week cohesiveness between its cells and the surrounding cells (i.e. small $T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)$). Thus, we define the structural consistency of a part filter $\boldsymbol{F}_i$ by

$$U(\boldsymbol{F}_i) = \frac{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i)}{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) + T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}. \tag{11}$$

With Eq. (11), the structural consistency of the part filter configuration Λ is computed by

$$S_C(\Lambda) = \sum_{\boldsymbol{F}_i \in \Lambda} U(\boldsymbol{F}_i). \tag{12}$$

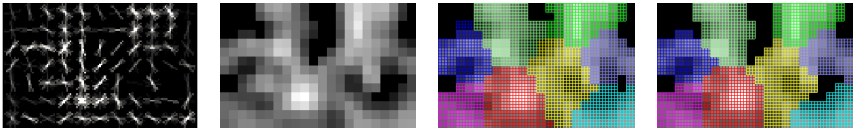
Optimization. Since there are $(2H_0 \times 2W_0)^{N_c}$ possible part filter configurations, it is computationally expensive to obtain the optimal part-filter configuration Λ^* by exhaustive search. Thus, we solve the optimization problem by first choosing an initial part-filter configuration Λ_0 which has a high structural consistency $S_C(\Lambda_0)$ and then refining Λ_0 to find a local optimum for the objective function in Eq. (5). We obtain Λ_0 by partitioning the enlarged root filter $\boldsymbol{F}_0'$ into N_c part filters such that their total structural consistency is maximized:

$$\Lambda_0 = \arg\max_{\Lambda} S_C(\Lambda) = \arg\max_{\Lambda} \sum_{i=1}^{N_c} \frac{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i)}{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) + T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}. \tag{13}$$

As $\frac{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i)}{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) + T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)} = 1 - \frac{T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) + T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}$, Eq. (13) is equivalent to

$$\Lambda_0 = \arg\min_{\Lambda} \sum_{i=1}^{N_c} \frac{T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}{T_W(\boldsymbol{F}_i, \boldsymbol{F}_i) + T_B(\boldsymbol{F}_i, \bar{\boldsymbol{F}}_i)}. \tag{14}$$

The objective function in Eq. (14) has the same form as the K-way normalized cuts criterion and an approximate solution can be obtained by the multiclass spectral clustering method [36].



(a) (b) (c) (d)

Figure 3: Our part discovery method. (a) The root filter of the right facing sub-category of bicycle. (b) The average response map of cells. (c) The initial part-filter configuration Λ_0. The black grids denote the cells which are removed from the enlarged root filter. (d) The final part-filter configuration after refinement. For clarity, $\tau = 0$ is used in this example to obtain the part-filter configuration without overlap.

Table 1: Average precision (AP) for 10 object categories in the PASCAL VOC2007 dataset with different λ.

	aero	bike	cat	horse	sheep	sofa	bus	dog	bottle	cow
$\lambda = 0.6$	33.9	59.7	20.5	58.8	**24.5**	38.5	52.9	12.5	**27.4**	**27.3**
$\lambda = 0.7$	33.1	60.1	22.3	**59.6**	22.6	38.4	51.0	12.3	**27.4**	24.9
$\lambda = 0.8$	34.0	**60.2**	**23.6**	59.0	24.2	37.1	53.3	**12.8**	**27.4**	26.8
$\lambda = 0.9$	34.5	58.9	22.8	57.0	22.1	38.7	**53.6**	12.6	27.1	26.3
$\lambda = 1.0$	**34.9**	59.7	22.5	57.9	23.2	**38.9**	51.9	12.6	26.8	25.2

Generally, object examples are not exactly rectangular and some cells in the enlarged root filter $\boldsymbol{F}'_0$ may correspond to the background. These background cells possibly cause small isolated part filters corresponding to the background to be included in the initial part filter configuration Λ_0. In our method, before partitioning $\boldsymbol{F}'_0$, we remove 10% of the cells near the boundary of $\boldsymbol{F}'_0$ with low average responses over object examples in D_c. The average response of a cell C_k is computed by

$$R_C(C_k) = \frac{\sum_{j=1}^{K} \boldsymbol{F}'_0(C_k) \cdot \boldsymbol{f}_j(C_k)}{K}, \tag{15}$$

where $\boldsymbol{F}'_0(C_k)$ is the weight vector of C_k and $\boldsymbol{f}_j(C_k)$ is the corresponding HOG feature. In the refinement step, we iteratively update the current part-filter configuration. Possible update operations include transferring one cell from a part filter to another, removing one cell from a part filter and adding one cell to a part filter. We only choose update operations that satisfy the connectivity constraint on every part filter and the overlapping constraint on the part filter configuration. If the new part filter configuration obtained by an update operation has a higher fitness value defined by Eq. (5), we replace the current part-filter configuration with the new one. The refinement process continues until there is no update operation that can lead to a better part-filter configuration. Figure 3 illustrates the process of our part discovery method.

4 Experiments

We test our approach on two benchmark datasets, PASCAL VOC2007 [9] and VOC2010 [10] datasets, which are commonly used by most recent work on generic object detection. Following the protocol of comp3 [9], we use the train and validation subsets for training

Table 2: Performance comparison using Average Precision (AP) on 20 object categories in the PASCAL VOC2007 dataset.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
DPM-V5	32.4	57.7	10.7	15.7	25.3	51.3	54.2	17.9	21.0	24.0	25.7
And-Or	**35.3**	**60.2**	11.0	16.6	**29.5**	53.0	**57.1**	23.0	**22.9**	**27.7**	28.6
Ours	34.0	**60.2**	**12.2**	**18.3**	27.4	**53.3**	55.9	**23.6**	22.6	26.8	**30.8**

	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
DPM-V5	11.6	55.6	47.5	**43.5**	14.5	22.6	34.2	44.2	41.3	32.5
And-Or	**13.1**	58.9	**49.9**	41.4	**16.0**	22.4	**37.2**	**48.5**	42.4	**34.7**
Ours	12.8	**59.0**	49.5	42.6	15.4	**24.2**	37.1	44.8	**43.4**	**34.7**

Table 3: Performance comparison using Average Precision (AP) on 20 object categories in the PASCAL VOC2010 dataset.

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
DPM-V5	42.9	47.2	10.3	11.1	26.3	48.4	40.2	22.9	17.0	22.9	10.2
And-Or	44.6	48.5	**10.8**	**12.9**	26.3	47.5	41.6	21.6	17.3	23.6	11.5
Ours	**45.9**	**50.7**	10.4	11.2	**28.7**	**50.5**	**44.2**	**24.2**	**17.4**	**24.0**	**13.7**

	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
DPM-V5	19.9	**41.5**	44.0	**41.0**	7.6	28.3	18.2	39.0	32.9	28.6
And-Or	**22.9**	40.9	45.3	37.9	**9.6**	**30.4**	**25.3**	39.0	31.2	29.4
Ours	17.6	40.2	**45.7**	38.8	8.3	29.3	20.2	**41.0**	**35.6**	**30.0**

and the test subset for testing. Average precision (AP) is used as evaluation measure for each object category and mean AP is computed over all object categories. We compare our approach with two most related approaches, DPM-V5 [17] and And-Or [31], which adopt the greedy search approach and the And-Or tree model respectively to initialize part filters for DPM.

Implementation details. In our experiments, $\lambda = 0.8$ and $\sigma = 0.5$ are used for all object categories. For the overlap threshold τ, we do not set it to a fixed value. Instead, we restrict each part filter to shrink or expand within a 2-cell width band along the part filter boundary in the refinement step. We find this dynamic setting of τ works well in practice. To make a fair comparison, we use the same setting as DPM-V5 for part initialization: the number of part filters for each component is set to 8 and the deformation parameters of each part filter are set to $[0.1, 0, 0.1, 0]$. After model parameters are initialized, the full model is trained in the same way as in DPM-V5. Although the shapes of parts in our model are not rectangular, for simplicity we still represent each part by a rectangular filter with a corresponding mask to disable the unused cells. As our model has roughly the same number of cells in all part filters as the model obtained by DPM-V5, the computational complexity of our method is similar to that of DPM-V5.

Selection of parameter λ. We test our part discovery approach on 10 object categories selected from PASCAL VOC2007 with λ ranging from 0.6 to 1.0. The results are listed in Table 1. It can be seen that a bad choice of λ may lead to a significant performance decrease for some object categories. Take the cat category for example. The performance of $\lambda = 0.6$ decreases by 13% compared to that of $\lambda = 0.8$. As shown in Table 1, $\lambda = 0.8$ works well for most of the object categories. We thus use this setting of λ to test our approach on the PASCAL VOC2007 and VOC2010 datasets. Cross-validation can also be used to select an

appropriate λ for each object category.

Experimental results. Detection results of DPM-V5, And-Or and our approach on the PASCAL VOC2007 and VOC2010 datasets are given in Tables 2 and 3 respectively. All the results are obtained without any post-processing, like bounding-box prediction or context rescoring. Our approach outperforms DPM-V5 for 19 and 17 out of 20 classes in PASCAL VOC2007 and VOC2010 respectively, which demonstrates the advantage of the use of non-rectangular part filters. Our approach achieves significant performance improvements for a few object categories, like bicycle, dining table and tv monitor, which have relatively stable structures. Overall, our approach is comparable to And-Or and performs slightly better than And-Or on PASCAL VOC2010. Besides the use of non-rectangular part filters, another difference between our approach and And-Or is that our approach uses 8 part filters for all object categories, while And-Or can automatically determines the number of part filters for each object category. We will study part number selection for DPM in future work to check if it can further improve the performance. Figure 4 shows some detection examples of our approach for the 20 object categories in PASCAL VOC2007.

5 Conclusions

We present a data-driven approach to discover parts for DPM. Different from most DPM based methods which use a specified number of rectangular parts, our approach is capable of discovering non-rectangular parts which can better fit object structures. Our approach can be efficiently implemented by solving a K-way normalized cuts problem followed by local refinement. The effectiveness of the proposed approach is validated on PASCAL VOC2007 and VOC2010 datasets. In future work, we will explore how to automatically select the part number for each object category.

Acknowledgement

This work is supported in part by Nanyang Assistant Professorship SUG M58040015.

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *European Conference on Computer Vision (ECCV)*, 2012.

[2] Dubout C. and Fleuret F. Exact acceleration of linear object detection. In *European Conference on Computer Vision (ECCV)*, 2012.

[3] G. Chen, Y. Ding, J. Xiao, and T. Han. Detection evolution with multi-order contextual co-occurrence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[4] Y. Chen, L. Zhu, and A. Yuille. Active mask hierarchies for object detection. In *European Conference on Computer Vision (ECCV)*, 2010.

[5] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

Figure 4: Detection examples of our approach for the 20 object categories in the PASCAL VOC2007 dataset.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[7] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] S. Divvala, A. Efros, and M. Hebert. How important are "deformable parts" in the deformable parts model. In *European Conference on Computer Vision (ECCV)*, 2012.

[9] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[10] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2010 (voc2010) results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010.

[11] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005.

[12] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 2010.

[14] S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[15] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 1973.

[16] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[17] R. Girshick, P. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/, 2012.

[18] C. Gu, P. Arbelaez, Y. Lin, K. Yu, and J. Malik. Multi-component models for object detection. In *European Conference on Computer Vision (ECCV)*, 2012.

[19] F. Khan, R. Anwer, J. Weijer, A. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[20] I. Kokkinos. Shufflets: Shared mid-level parts for fast object detection. In *International Conference on Computer Vision (ICCV)*, 2013.

[21] T. Lan, M. Raptis, L. Sigal, and G. Mori. From subcategory to visual components: A multi-level framework for object detection. In *International Conference on Computer Vision (ICCV)*, 2013.

[22] D. Levi, S. Silberstein, and A. Bar-Hillel. Fast multiple-part based object detection using kd-ferns. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[23] R. Mottaghi. Augmenting deformable part models with irregular-shaped object paches. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[24] R. Mottaghi, A. Ranganathan, and A. Yuille. A compositional approach to learning part-based models for objects. In *International Conference on Computer Vision (ICCV)*, 2011.

[25] O. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman. The truth about cats and dogs. In *International Conference on Computer Vision (ICCV)*, 2011.

[26] M. Pedersoli, A. Vedaldi, and J. GonzÍd'lez. A coarse-to-fine approach for fast deformable object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[27] X. Ren and D. Ramanan. Histograms of sparse codes for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[28] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele. Discriminative structure learning of hierarchical representations for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[29] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent crfs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[30] H. Song, S. Zickler, T. Althoff, R. Girshick, M. Fritz, C. Geyer, P. Felzenszwalb, and T. Darrell. Sparselet models for efficient multiclass object detection. In *European Conference on Computer Vision (ECCV)*, 2012.

[31] X. Song, T. Wu, Y. Jia, and S. Zhu. Discriminatively trained and-or models for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[32] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[33] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[34] W. Yang, Wang Y., and G. Mori. Recognizing human actions from still images with latent poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[35] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[36] S. Yu and J. Shi. Multiclass spectral clustering. In *International Conference on Computer Vision (ICCV)*, 2003.

[37] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[38] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. In *International Conference on Computer Vision (ICCV)*, 2010.