Learning Hybrid Part Filters for Scene Recognition

Yingbin Zheng, Yu-Gang Jiang, Xiangyang Xue

School of Computer Science, Fudan University, Shanghai, China
{ybzh, ygj, xyxue}@fudan.edu.cn

Abstract. This paper introduces a new image representation for scene recognition, where an image is described based on the response maps of object part filters. The part filters are learned from existing datasets with object location annotations, using deformable part-based models trained by latent SVM [1]. Since different objects may contain similar parts, we describe a method that uses a semantic hierarchy to automatically determine and merge filters shared by multiple objects. The merged *hybrid* filters are then applied to new images. Our proposed representation, called Hybrid-Parts, is generated by pooling the response maps of the hybrid filters. Contrast to previous scene recognition approaches that adopted object-level detections as feature inputs, we harness filter responses of object parts, which enable a richer and finer-grained representation. The use of the hybrid filters is important towards a more compact representation, compared to directly using all the original part filters. Through extensive experiments on several scene recognition benchmarks, we demonstrate that Hybrid-Parts outperforms recent state-of-the-arts, and combining it with standard low-level features such as the GIST descriptor can lead to further improvements.

1 Introduction

The construction of good image representations is of fundamental importance in many computer vision problems. Great progress has been achieved in the past years with the invention of local invariant descriptors like SIFT [2] and representations such as the bag-of-features [3] and its augmented version, the spatial pyramids [4]. Although promising results were shown from using these low-level representations, more recent research suggests that mid-level representations, where each dimension is associated with a semantic meaning, are more flexible and powerful for visual recognition [5–7]. The mid-level semantics, a.k.a. attributes (e.g., objects and scenes), can be learned offline, using existing datasets or open data on the Web.

In this paper, we propose a new image representation called Hybrid-Parts, which is built upon a large set of object part filters generated automatically by deformable part-based models (DPM) [1]. Starting from publicly available image sets annotated with object-level bounding boxes, DPM is employed to learn local part filters using Felzenszwalb's latent SVM formulation [1]. Since

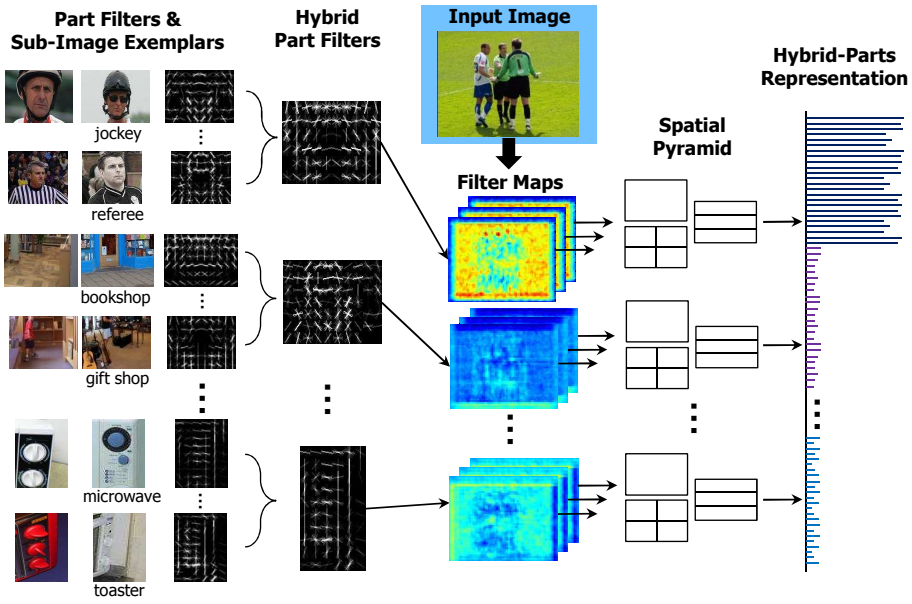

Fig. 1. An illustration of Hybrid-Parts representation. A large set of object part filters are firstly learned from existing object labels (object class names are shown below the part filters and sub-image exemplars). Similar part filters are then merged to generate hybrid filters, which are applied to an input image at multiple scales. The response maps are consolidated with a spatial pyramid of three levels to produce the final Hybrid-Parts representation, which simply concatenates the max response values of each hybrid filter in each image scale and spatial grid ($\#\text{filters} \times \#\text{scales} \times (1+2\times 2+3\times 1)$ dimensions).

different objects may share similar parts, we further introduce a method to determine and merge such common patterns, leading to a smaller set of *hybrid* part filters. The discovery of the common parts is done by a simple variant of random ferns [8], operated on top of a semantic hierarchy. Finally, given an input image, filter responses of the hybrid parts are consolidated into a semantic level representation, with each dimension indicating the likelihood of seeing an (hybrid) object part in the image.

Figure 1 illustrates our approach. As can be seen on the left side of the figure, some part filters show visually very similar patterns, e.g., the "person head" part from object classes "jockey" and "referee". This confirms the fact that the part filter set learned from multiple objects contains a significant amount of redundant information. In Hybrid-Parts, a more compact representation can be achieved with a smaller number of hybrid filters. Note that it is important to merge the filters before generating the response maps, because the other option, i.e., fusing the response maps of multiple part filters, apparently demands more computational workloads.

The idea of constructing mid-level image representations has been taken in several previous works, e.g., the Object Bank by Li *et al.* [7]. Compared to these approaches using image-level object or scene classification outputs as features, our Hybrid-Parts representation is based on the detections of local object parts. The utilization of finer-grained part-based semantics makes Hybrid-Parts more discriminative than the existing works. Our approach shows very competitive results on popular scene recognition benchmarks.

The remainder of this paper is organized as follows. Section 2 discusses related works and Section 3 briefly reviews the deformable part-based models. Our proposed Hybrid-Parts representation is described in Section 4 and extensively evaluated in scene recognition experiments in Section 5. Finally, Section 6 concludes this paper.

2 Related Works

Numerous efforts have been devoted to the design of effective image representations. In this section, we mainly discuss literatures on mid-level representation, which are more related to this work.

In [9], Farhadi *et al.* proposed to describe objects using mid-level semantics, namely attributes, including shape (e.g., cylindrical), part (e.g., head), and material (e.g., glass). In addition to textually describing objects, the attributes were also applied to learn new object classes with few or no examples. This method was further extended in [10] where the outputs of localized object detectors were used as attributes. Lampert *et al.* [11], Rohrbach *et al.* [12], and Yu *et al.* [13] explored a similar pipeline that selects and transfers pre-computed attributes for learning unseen objects. Wang and Forsyth [14] used multiple instance learning to learn attributes and object classes jointly and showed that an iterative refinement procedure leads to substantial performance improvement. The joint modeling of objects and attributes was also attempted by Wang and Mori [15], using an undirected graphical model. Torresani *et al.* [6] proposed Classemes as a descriptor for object recognition, which is generated by the classification outputs of 2659 semantic concepts (each can be viewed as an attribute). In addition to object recognition and detection, Hauptmann *et al.* [16] used a large set of semantic concept classifiers for improved video retrieval. Liu *et al.* [17] proposed to use attributes for action recognition. Berg *et al.* [18] presented an approach for automatic attribute discovery and modeling by mining noisy Web data. The learned attributes were applied to product image search.

Several works have studied mid-level representations in scene recognition [5, 7, 19]. In [5], Vogel and Schiele proposed to detect a set of visual concepts locally over image regions and the images were represented by the frequency of the detected local concepts. Parikh and Grauman [19] further proposed relative attributes to model the relative strength of seeing an attribute, rather than using hard binary occurrence labels. Another work more related to ours is the Object Bank [7], where hundreds of object detectors were trained with off-the-

shelf approaches like the DPM [1]. The outputs of object detectors were then consolidated to form the final image representation.

These existing approaches construct mid-level representations using image-level attribute predictions [6, 9, 11–13, 15, 16], region-based classification [5], or local object detection [7, 10]. Hybrid-Parts is similar to the latter ones [7, 10] in the sense that we also rely on object detection. However, instead of using the final object detection outputs like [7], we adopt the filter responses of object parts. Detectors of object parts were also used in [10] to enhance whole object detection, where the object parts were pre-defined and manually annotated. In contrast, the part filters in our approach are automatically learned by DPM, which requires much less annotation efforts. Very recently, DPM was employed by Pandey and Lazebnik [20] *directly* for scene recognition and the learned *scene* parts from DPM may correspond to recurring elements or objects in a scene class. Our approach is fundamentally different from [20] in its design since we use object parts learned from external data. Moreover, as will be shown in the experiments, Hybrid-Parts outperforms these state-of-the-art methods.

3 Deformable Part-Based Models

This section briefly introduces the DPM framework proposed in [1], which is adopted in this work to learn the basic object part filters.

First, Histogram of Oriented Gradients (HOG) [21] is computed to represent an image in a multi-scale feature pyramid. Using multi-scale features is a standard setup for scale-invariant object detection. For the parameters in the HOG feature, e.g., the number of pyramid levels and orientations, we follow the original settings of [1][1].

DPM uses a star model to characterize an object, which consists of a coarse root filter that roughly covers the entire object and several part filters that cover small object parts. A filter (rectangular template) is defined by an array of weight vectors. Let $\mathbf{w}$ be a long vector obtained by concatenating the weight vectors of a filter in row-major order. Let H be the HOG feature pyramid and p specify the level of the pyramid and a candidate position. The score of the filter vector $\mathbf{w}$ at p can be computed as the dot product of $\mathbf{w}$ and $\phi(H, p)$, where $\phi(H, p)$ returns a vector that concatenates the feature values in a subwindow of H with top-left corner at p, also in row-major order.

Denote $z = (p_r, p_1, \ldots, p_n)$ as a hypothesis of an object which specifies the positions of the root filter (p_r) and the part filters ($p_i, i = 1, \ldots, n$). The score of the object hypothesis z is defined as

$$s(z) = \mathbf{w}_r \cdot \phi(H, p_r) + \sum_{i=1}^{n} \mathbf{w}_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} \mathbf{d}_i \cdot \phi_d(dx_i, dy_i) + b, \qquad (1)$$

where $\mathbf{w}_r$ is the root filter vector, $\mathbf{w}_i$ is the vector of the ith part filter, $\phi_d(dx_i, dy_i)$ computes the displacement of the ith part relative to its anchor position, $\mathbf{d}_i$ is a

[1] Codes from the authors of [1] are at `http://www.cs.brown.edu/~pff/latent/`.

deformation cost vector for the ith part, and b is a bias term to make the output scores of multiple models comparable.

A latent SVM formulation is used to learn the DPM parameters, denoted by a single vector $\beta = (\mathbf{w}_r, \mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{d}_1, \ldots, \mathbf{d}_n, b)$. In the latent SVM, an example X is scored by a function of the following form:

$$f_\beta(X) = \max_z \beta \cdot \mathbf{\Phi}(X, z), \tag{2}$$

where z are latent values (for a hypothesis); $\mathbf{\Phi}(X, z)$ is a concatenation of sub-windows from a feature pyramid of X according to the latent values in z. Model training is conducted by alternatively executing two steps. First, the parameters in β are fixed and optimization is performed for the latent values over all the positive examples. After that, the latent values are fixed and β is optimized using an objective function similar to standard SVM. The training procedure also includes several important settings such as the selection of suitable negative samples and the initialization of model parameters. We refer the reader to [1] for more details.

4 Hybrid-Parts

We now elaborate the construction of Hybrid-Parts. We start by introducing the object labels used for training the part filters, followed by a method that determines and merges the filters shared by multiple objects, i.e., the generation of the hybrid filters.

4.1 Object Labels

The selection of a suitable set of object labels is important to the success of Hybrid-Parts. Ideally we want a comprehensive collection of objects so that we can find some of them under any scene setting. As pointed out by [7], the objects used here should not be limited to the traditional ones like "washer" and "banana". Some generalized object classes like "volcano" and "bookshop" may also be included. With these in mind, we selected the object labels from ImageNet [22]. Images in ImageNet are organized by the WordNet hierarchy [23], where each class is called a synset. Local object-level bounding box annotations[2] are available for around 3,000 synsets (object classes). This set was further filtered by two criteria. First, to make sure the DPM has adequate training samples, classes with less than 100 example images were removed. Second, among the remaining ones, we further removed several higher level classes in the WordNet hierarchy and only kept the leaf nodes, as examples of the high-level classes may be visually too diverse to train a good DPM. This leads to a final set of 580 object classes, covering a wide range of entities in the visual world, such as appliance (e.g., "washer"), fruit (e.g., "banana"), fabric (e.g., "paper towel") and structure (e.g., "bookshop"). Figure 2 shows several example images with object bounding box annotations.

[2] Download link: `http://www.image-net.org/download-bboxes`.

Fig. 2. Example training images with object bounding box annotations.

4.2 Generating Hybrid Filters

DPM is trained for each of the object classes, producing one root filter and six part filters. The part filters can be directly applied to a test image and generate a mid-level representation by pooling the corresponding response maps on a spatial pyramid. However, many objects may share similar parts, resulting in a significant amount of redundant information in the part filter set. In this section, we introduce a simple method that merges similar part filters to generate hybrid parts. The use of hybrid parts not only largely reduces the dimensionality of the final representation, but also offers a semantically more balanced representation, since some frequently "repetitive" parts in the original part filter set (e.g., the *head part* of person-related objects) may dominant the infrequent ones.

Central to the generation of the hybrid parts is to identify the sets of similar part filters. Intuitively the similarity of the part filters may be computed by comparing their corresponding weight vectors $\mathbf{w}$. But we have found this way not very effective (cf. Section 5.2 for evaluations). We therefore pursue an exemplar-based approach that measures part similarity based on exemplar sub-images. To find the exemplar sub-images, we run DPM object detection back on the training images, which produces bounding boxes of both the entire object and its deformable parts for each detection. This way we can easily generate a set of exemplars for each part filter. Figure 3 shows some examples.

With the exemplar sets, the similarity of object parts can be estimated based on the proximities of their corresponding exemplars in feature space. Instead of exhaustively computing proximities among all the exemplar sets, here we propose to impose a semantic constraint which only allows to merge the filters from objects that are close in semantic space. In other words, we are more interested

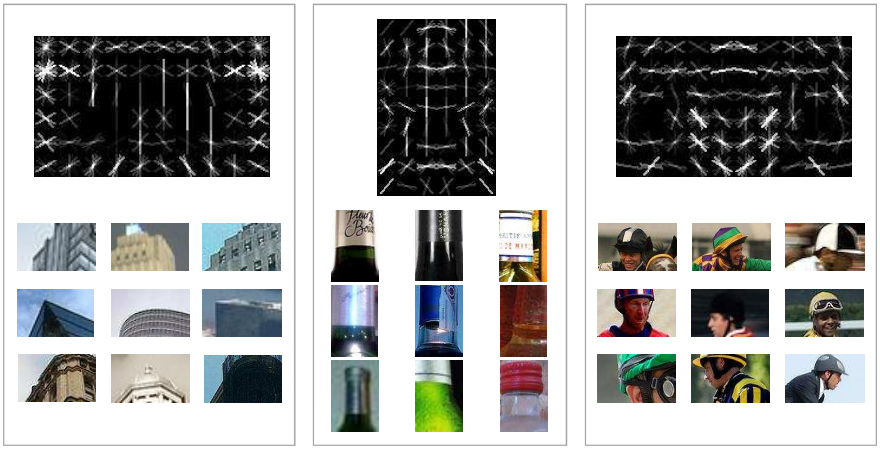

Fig. 3. Three object part filters, each with a few sub-image exemplars obtained from the ImageNet training set.

in the relations among the parts of objects like "dog" and "cat", rather than that of "dog" and "table". To this end, we adopt the hierarchical structure of WordNet to merge the filters in a bottom-up manner. Lowest common ancestors (LCA) of the leaf-node objects are picked (in total there are 145 LCA nodes), and hybrid filters are generated for each LCA node by merging similar part filters from its child nodes.

Formally, let $\mathcal{S}$ be an exemplar set of an LCA node $\mathcal{O}$, which is simply the pool of all the exemplars from the child nodes of $\mathcal{O}$. We employ a variant of the random ferns [8] (i.e., a simplified random forest [24]) to determine similar parts from the child nodes, by quickly estimating the similarities of the corresponding exemplars in $\mathcal{S}$. HOG features are computed for the exemplar sub-images using the same number of grids, orientations and pyramid levels. Denote the HOG feature of an exemplar x as $\mathbf{v}_x$. $\mathcal{S}$ is split into two sets by a linear random projection (with parameters $\mathbf{p}$ and t) in the HOG feature space:

$$x \in \begin{cases} \mathcal{S}_l & \text{if } \mathbf{p} \cdot \mathbf{v}_x \leq t, \\ \mathcal{S}_r & \text{otherwise.} \end{cases} \tag{3}$$

To make sure the split is "balanced", several candidates for $\mathbf{p}$ and t are generated, and we choose the one that maximizes the information gain below:

$$\Delta E = -\sum_{j=1}^{J} \left(\frac{|\mathcal{S}_l^j|}{|\mathcal{S}^j|} E(\mathcal{S}_l^j) + \frac{|\mathcal{S}_r^j|}{|\mathcal{S}^j|} E(\mathcal{S}_r^j) \right), \tag{4}$$

where J is the number of the child-node classes under the LCA node $\mathcal{O}$; $\mathcal{S}^j$ is the exemplar set of the jth child-node object class; $\mathcal{S}_l^j$ and $\mathcal{S}_r^j$ are the split sets of $\mathcal{S}^j$ by the random projection. The function $E(\cdot)$ computes the entropy of a

set, defined as $E(\mathcal{S}_l^j) = -\sum_{i=1}^{n} P(a_i^j|\mathcal{S}_l^j)\log_2(P(a_i^j|\mathcal{S}_l^j))$, where a_i^j indicates the ith part of the jth child-node class, n is the total number of parts of the class, and $P(a_i^j|\mathcal{S}_l^j)$ is the proportion of exemplars in $\mathcal{S}_l^j$ belonging to part a_i^j.

Based on a random fern defined by a selected random projection, we measure the part similarities in a part set $\mathcal{A} = \{a_k\}_{k=1}^{K}$ as

$$\mathcal{M}(\mathcal{A}) = \frac{1}{2}\left(\prod_k P(a_k|\mathcal{S}_l^{c(a_k)}) + \prod_k P(a_k|\mathcal{S}_r^{c(a_k)})\right), \tag{5}$$

where $c(a_k)$ returns the index of the child-node class to which a_k belongs. As can be understood from the above definition, the function value of $\mathcal{M}(\cdot)$ is maximized when the exemplars of the parts in $\mathcal{A}$ are all on the same side of the random fern projection. Therefore $\mathcal{M}(\cdot)$ reflects the similarities of the parts in $\mathcal{A}$. But one random fern is obviously not enough. In our experiments, 50 random ferns are used and the mean value of $\mathcal{M}(\cdot)$ is adopted to measure the overall part similarities of a part set.

Now that we have an efficient way to estimate part proximities in a set. Given an LCA node, its candidate part sets are formed by choosing at most one part from each child-node class. Each candidate part set is then evaluated by Equation (5) and finally the sets with high $\mathcal{M}(\cdot)$ values are selected. Since the part similarity is estimated by the grid-based HOG feature of the corresponding exemplars, the part filters in a selected set not only share similar appearance but also have similar spatial configurations. Therefore we simply align the part filters in each set at their top-left corner and average them to generate a hybrid filter. Given a test image, the Hybrid-Parts representation is generated by pooling the response scores of the hybrid filters over a spatial pyramid, as illustrated in Figure 1. Generating hybrid filters for a selected part set is preferred compared to averaging the response maps of all the filters in the set, since we only need to compute one response map for the former, which is apparently much more efficient.

5 Experiments

We evaluate the proposed Hybrid-Parts representation on three popular datasets: MIT Indoor Scene [25], Fifteen Natural Scene [4], and UIUC Sports [26]. The MIT Indoor Scene contains 15,620 images labeled over 67 classes. We adopt the official train/test split of [25] to use 80 images per class for training and 20 for testing. The Fifteen Scene dataset [4] contains 15 classes, with 200–400 images per class. We follow the standard setup of this dataset to randomly select 100 images from each class for training and use the rest for testing. The UIUC Sports has 8 complex sports scene classes (e.g., *badminton* and *croquet*). Each class has around 200 images. Following [26], we use 70 randomly drawn images per class for training and 60 for testing. For both Fifteen Scene and UIUC Sports, we perform ten random selections of the training and testing images and report the mean accuracy over the ten runs. Throughout the experiments, we use the highly efficient multi-class linear SVMs for scene classification.

Table 1. Classification accuracy for MIT Indoor Scene dataset. We compare Hybrid-Parts with several baselines and state-of-the-art approaches. Hybrid-Parts offers the best single representation performance, and fusing it with a few baseline descriptors leads to a significant performance gain.

	Approach	Accuracy
Baselines	GIST [27]	22.0%
	GIST-color [27]	29.7%
	HOG [21]	22.8%
	Spatial Pyramid (SP) [4]	34.4%
	GIST-color + SP	38.5%
State of the arts	ROI + GIST [25]	26.5%
	CENTRIST [28]	36.9%
	Object Bank [7]	37.6%
	Classemes [6]	38.6%
	DPM [20]	30.4%
	DPM + GIST-color + SP [20]	43.1%
Our results	Hybrid-Parts	39.8%
	Hybrid-Parts + GIST-color + SP	47.2%

5.1 Results and Comparison

We first evaluate the performance of Hybrid-Parts and compare it with several state-of-the-art approaches. There are a few parameters in Hybrid-Parts, including the number of image scales for filtering, spatial pyramid levels, and the number of hybrid filters generated per LCA node (based on the function value of $\mathcal{M}$; see Section 4.2). In this set of experiments, we use 3 image scales, 3 spatial pyramid levels (1×1, 2×2, and 3×1), and generate 12 hybrid filters per LCA node (145×12=1740 hybrid filters in total). The effect of these parameters will be evaluated in the next subsection.

MIT Indoor. Table 1 summarizes the average classification accuracies of all the classes in the MIT Indoor dataset. We compare Hybrid-Parts with several baselines and state-of-the-art approaches. Among them, the first group contains several well-known baseline descriptors, including GIST [27], GIST-color (concatenation of GIST descriptors of RGB color channels), HOG [21], and Spatial Pyramid (SP) [4]. These baseline performance numbers are from Pandey and Lazebnik [20]. From the table we see that Hybrid-Parts outperforms all the baseline descriptors as well as the state-of-the-art approaches shown in the middle part of Table 1 (except the fusion results of multiple classifiers, indicated by "+"), including a very recent holistic image descriptor CENTRIST [28], scene category modeling directly by DPM [20], and two mid-level representations Object Bank[3] [7] and Classemes [6]. The substantial performance gains over the state of the arts confirm the effectiveness of using object parts as the basic

[3] The Object Bank performance in Table 1 is from [7]. We also implemented the Object Bank representation using our object set (580 classes in total) and 3 spatial pyramid levels. The classification accuracy is 31.0%.

Table 2. Per-class classification rates (%) for MIT Indoor Scene dataset, using "Hybrid-Parts" (H) and "Hybrid-Parts+GIST-color+SP" (HGS). The classes are sorted in descending order of the "Hybrid-Parts" performance.

	H	HGS		H	HGS		H	HGS		H	HGS
cloister	85	85	inside bus	55	63	gym	35	75	kindergarden	26	46
corridor	82	84	bowling	54	63	bathroom	35	33	grocerystore	25	53
elevator	73	76	trainstation	52	59	clothingstore	35	43	operating room	25	38
studiomusic	71	60	kitchen	50	57	hairsalon	33	57	auditorium	24	55
greenhouse	67	67	library	47	38	laboratorywet	33	41	shoeshop	24	50
buffet	62	45	videostore	47	29	museum	33	50	livingroom	23	27
concert hall	62	57	bedroom	45	37	restaurant kitchen	33	50	waitingroom	22	17
closet	61	62	warehouse	44	50	fastfood restaurant	32	30	mall	20	40
pantry	59	55	computerroom	43	38	casino	31	65	toystore	19	22
church inside	58	56	dentaloffice	43	57	lobby	30	41	artstudio	18	24
florist	58	67	nursery	41	46	poolinside	30	44	winecellar	17	31
inside subway	58	55	subway	40	65	airport inside	29	25	locker room	14	43
movietheater	57	61	classroom	39	46	gameroom	29	58	restaurant	13	27
laundromat	56	48	meeting room	38	44	tv studio	29	80	jewelleryshop	10	22
garage	56	58	bookstore	36	29	children room	27	36	deli	6	20
prisoncell	56	73	dining room	36	42	bakery	27	20	office	4	24
stairscase	56	48	hospitalroom	36	28	bar	26	40			

elements for mid-level image feature construction. Another interesting observation is that the three mid-level representations (Object Bank, Classemes, and Hybrid-Parts) are better than all the other single feature/model approaches. This highlights the advantages of using mid-level representations to bridge the semantic gap between low-level features and high-level semantics.

We also study the fusion performance by combining the Hybrid-parts representation with two baseline descriptors GIST-color and SP. Results are also given in Table 1. The fusion is done by simply multiplying the softmax transformed prediction scores from multiple SVM classifiers as [20], and an image is assigned to the class with the highest score (confidence) after multiplication. As shown in the table, fusing Hybrid-Parts with two baseline descriptors leads to an accuracy of 47.2%, which is significantly higher than the fusion accuracy of [20] (43.1%). Table 2 lists the per-class performance of Hybrid-Parts and its fusion with GIST-color and SP.

Fifteen Scene. Results on the Fifteen Scene datasets are presented in Figure 4, where the accuracy numbers of GIST-color, SP and Classemes are based on our implementation. For GIST-color and Classemes, we use the software released by the authors of [27][4] and [6][5] respectively. For SP, we use a SIFT vocabulary of 500 codewords and 3 spatial pyramid levels. As can be seen from Figure 4-Left, Hybrid-Parts shows very competitive performance with a classification accuracy of 84.7%, which is significantly better than that of the other two mid-level representations Object Bank and Classemes. The approach LScSPM [29] is also a spatial pyramid representation, but is generated by Laplacian sparse coding of low-level local descriptors. We believe that Hybrid-Parts is also complementary to LScSPM since they focus on different levels of information, which is nevertheless difficult to verify since there is no public software for LScSPM.

UIUC Sports. We now report results on the UIUC Sports dataset [26], as summarized in Figure 5. Similar to the experiments on Fifteen Scene, the GIST-color, SP, and Classemes numbers are based on our implementation, partly using

[4] Available at `http://people.csail.mit.edu/torralba/code/spatialenvelope/`.
[5] Available at `http://www.cs.dartmouth.edu/~lorenzo/projects/classemes/`.

Approach	Accuracy
GIST-color [27]	69.5%
SP [4]	81.4%
SP-pLSA [30]	83.7%
CENTRIST [28]	83.9%
HIK [31]	84.1%
HG [32]	85.2%
LScSPM [29]	89.8%
Object Bank [7]	80.9%
Classemes [6]	80.6%
Hybrid-Parts	84.7%
Hybrid-Parts+GIST-color+SP	86.3%

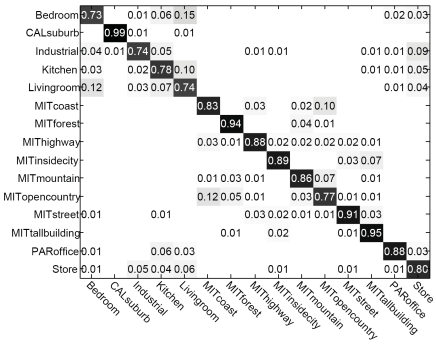

Fig. 4. Results for Fifteen Scene dataset. **Left.** Classification accuracy. **Right.** Confusion matrix of the results using Hybrid-Parts representation.

Approach	Accuracy
GIST-color [27]	70.7%
SP [4]	81.8%
Graphical Model [26]	73.4%
LDA [33]	66.0%
CENTRIST [28]	78.3%
HIK [31]	84.2%
LScSPM [29]	85.3%
Object Bank [7]	76.3%
Classemes [6]	84.2%
Hybrid-Parts	84.5%
Hybrid-Parts+GIST-color+SP	87.2%

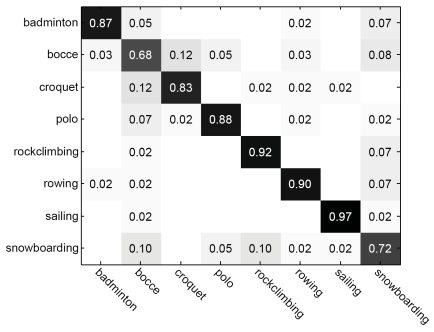

Fig. 5. Results for UIUC Sports dataset. **Left.** Classification accuracy. **Right.** Confusion matrix of the results using Hybrid-Parts representation.

the authors' codes. Again, Hybrid-Parts shows better results than the other two mid-level representations. Fusing it with the two simple baseline descriptors GIST-color and SP, we obtain an accuracy of 87.2%, which outperforms all the approaches under comparison, including the LScSPM.

5.2 Evaluation of Parameters and Alternative Implementation

In this subsection, we evaluate a few parameters and an alternative implementation for generating Hybrid-Parts. We report results on the MIT Indoor Scene dataset as it is larger and more challenging than the other two.

Number of Image Scales. Just like the general object detection task where a detector needs to run on multiple image scales to achieve scale-invariance, Hybrid-Parts also favors multi-scale image filtering. Figure 6-Left shows the classification accuracy of our Hybrid-Parts representation generated by performing part filtering on different number of image scales. We observe significant perfor-

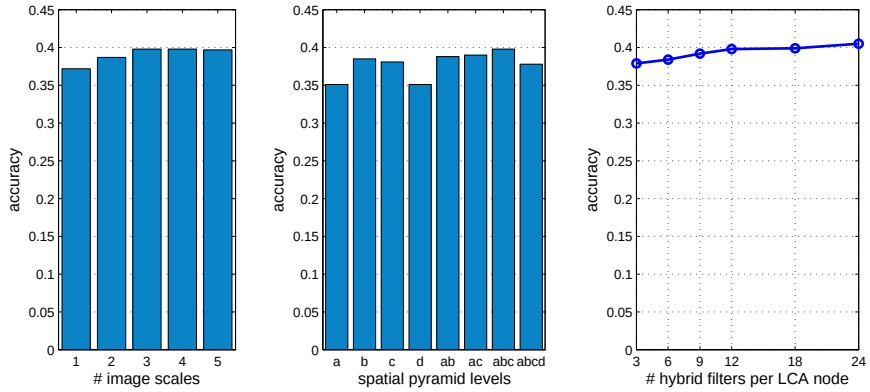

Fig. 6. Evaluation of Hybrid-Parts parameters on the MIT Indoor Scene dataset. **Left.** Classification accuracy w.r.t. the number of image scales for filtering. **Middle.** Classification accuracy with various spatial pyramid levels (a: 1×1, b: 2×2, c: 3×1, d: 4×4). **Right.** Classification accuracy w.r.t. the number of hybrid filters per LCA node. See texts for more details.

mance gains when the number of image scales increases from one to three. Using more scales does not improve the performance.

Spatial Pyramid Levels. We also evaluate the effect of spatial pyramid levels in Hybrid-Parts representation. As shown in Figure 6-Middle, the best results is obtained by using three spatial pyramid levels (1×1, 2×2, and 3×1). These coarse partitions of image regions are preferred compared to fine-grained partitions like 4×4 since the latter may incur significant mismatch problems when similar scene patterns appear in different regions. This is consistent with the observations from recent research on object/scene recognition [34].

Number of Hybrid Filters per LCA Node. Recall that the hybrid filters are generated from the selected part sets, based on the function value of $\mathcal{M}(\cdot)$ shown in Equation (5). In Figure 6-Right we plot the classification accuracy versus the number of hybrid filters per LCA node. Using only 3 hybrid filters per LCA node, we can already get an accuracy of 37.9%. Adding more filters further boosts the accuracy until 12 (39.8%), after which the performance tends to be saturate. We also found that this result is similar to the accuracy of using all the 3480 (580×6) original part filters of the child-node objects (40.4%), indicating that around half of the computation[6] can be saved by Hybrid-Parts with marginal performance drop.

Hybrid Filter Generation. Our last experiment evaluates an alternative method for selecting the object part sets in hybrid filter generation. As mentioned earlier, the similarities of part filters may be computed by directly comparing their corresponding weight vectors $\mathbf{w}$. This is achieved by firstly aligning the filters and then computing the similarity of the weights corresponding to the

[6] Hybrid-Parts has 1740 filters in total (145 LCA nodes, each with 12 hybrid filters).

overlapped areas. The part similarities computed in this way can be used to select the part filter sets and generate hybrid filters. With the same spatial pyramid levels, we obtain an accuracy of 38.6%, which is not as good as the exemplar-based method described in Section 4.2 (39.8%).

6 Conclusions

We have introduced Hybrid-Parts, a mid-level image representation based on the responses of object part filters. Through an extensive set of scene recognition experiments, we have shown that Hybrid-Parts is more effective than existing mid-level representations, generating very competitive results on popular benchmark datasets. Compared with directly employing all the original part filters of the leaf-node objects, Hybrid-Parts offers similar performance using a much smaller number of hybrid filters. Moreover, we also observed that Hybrid-Parts is complementary to popular baseline descriptors like the GIST, and combining them can further boost the recognition accuracy. For future work, we plan to incorporate more object classes into Hybrid-Parts, and also explore this representation in other tasks.

Acknowledgments. This work was supported in part by two STCSM's Programs (No. 10511500703 & No. 12XD1400900), a National 863 Program (No. 2011AA010604), and a National 973 Program (No. 2010CB327906).

References

1. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32** (2009) 1627–1645
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
3. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006) 2169–2178
5. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. IJCV **72** (2007) 133–157
6. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: ECCV. (2010)
7. Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification semantic feature sparsification. In: NIPS. (2010)
8. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: ICCV. (2007)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
10. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: CVPR. (2010)

11. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR. (2009)
12. Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., Schiele, B.: What helps where– and why? semantic relatedness for knowledge transfer. In: CVPR. (2010)
13. Yu, X., Aloimonos, Y.: Attribute-based transfer learning for object categorization with zero/one training example. In: ECCV. (2010)
14. Wang, G., Forsyth, D.: Joint learning of visual attributes, object classes and visual saliency. In: ICCV. (2009)
15. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV. (2010)
16. Hauptmann, A., Yan, R., Lin, W., Christel, M., Wactlar, H.: Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. TMM **9** (2007) 958–966
17. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR. (2011)
18. Berg, T., Berg, A., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: ECCV. (2010)
19. Parikh, D., Grauman, K.: Relative attributes. In: ICCV. (2011)
20. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV. (2011)
21. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
22. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
23. Miller, G.: Wordnet: a lexical database for English. Communications of the ACM **38** (1995) 39–41
24. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: CVPR. (2007)
25. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009)
26. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
27. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
28. Wu, J., Rehg, J.: CENTRIST: A visual descriptor for scene categorization. TPAMI **33** (2011) 1489–1501
29. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely – Laplacian sparse coding for image classification. In: CVPR. (2010)
30. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. TPAMI **30** (2008) 712–727
31. Wu, J., Rehg, J.M.: Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: ICCV. (2009)
32. Zhou, X., Cui, N., Li, Z., Liang, F., Huang, T.: Hierarchical gaussianization for image classification. In: ICCV. (2009)
33. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: CVPR. (2009)
34. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. TMM **12** (2010) 42–53