Randomized Max-Margin Compositions for Visual Recognition

Angela Eigenstetter[1] Masato Takami[1,2] Björn Ommer[1]
[1]Heidelberg Collaboratory for Image Processing & IWR, University of Heidelberg, Germany
[2]Robert Bosch GmbH, Corporate Research, Hildesheim, Germany
{aeigenst,mtakami,bommer}@iwr.uni-heidelberg.de

Abstract

A main theme in object detection are currently discriminative part-based models. The powerful model that combines all parts is then typically only feasible for few constituents, which are in turn iteratively trained to make them as strong as possible. We follow the opposite strategy by randomly sampling a large number of instance specific part classifiers. Due to their number, we cannot directly train a powerful classifier to combine all parts. Therefore, we randomly group them into fewer, overlapping compositions that are trained using a maximum-margin approach. In contrast to the common rationale of compositional approaches, we do not aim for semantically meaningful ensembles. Rather we seek randomized compositions that are discriminative and generalize over all instances of a category. Our approach not only localizes objects in cluttered scenes, but also explains them by parsing with compositions and their constituent parts.

We conducted experiments on PASCAL VOC07, on the VOC10 evaluation server, and on the MITIndoor scene dataset. To the best of our knowledge, our randomized max-margin compositions (RM^{2}C) are the currently best performing single class object detector using only HOG features. Moreover, the individual contributions of compositions and their parts are evaluated in separate experiments that demonstrate their potential.

1. Introduction

Discriminative part-based models currently constitute one of the most popular and powerful paradigms for the challenging problem of category-level object detection such as PASCAL VOC [9]. The underlying rationale is to use a small number of informative parts and combine them with a powerful discriminative approach such as the deformable part model [12] based on their appearance and location. This framework typically restricts such discriminative methods as [12, 32] to only few parts, as opposed to weaker spatial models such as bag-of-features [6], Hough voting [21], or generative methods such as [11, 18, 28]. Another paradigm is to use extra manual part labellings as in [3], which is only feasible for a limited number of parts. In contrast we aim for a large number of specific but weak parts (on the order of 1000 per category) that are automatically learned on comparably few training samples (around 100 positives per class) without requiring extra annotations. Each of these parts is trained on only a small region of a single positive sample against negatives. In contrast to other part-based methods such as [8, 16, 22, 29], we compensate for the weakness of specialized, local, and frail parts by grouping them into stronger compositions that exhibit im-

Figure 1. Object detection and parsing with randomized max-margin compositions (RM^{2}C). The discriminative approach not only detects object, but also activates compositions according to the classification function $g(F(I))$. Compositions in turn activate parts i (we plot the corresponding positive training patch x_p) by weighting them according to the decision function f_k.

In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14), ©2014 IEEE

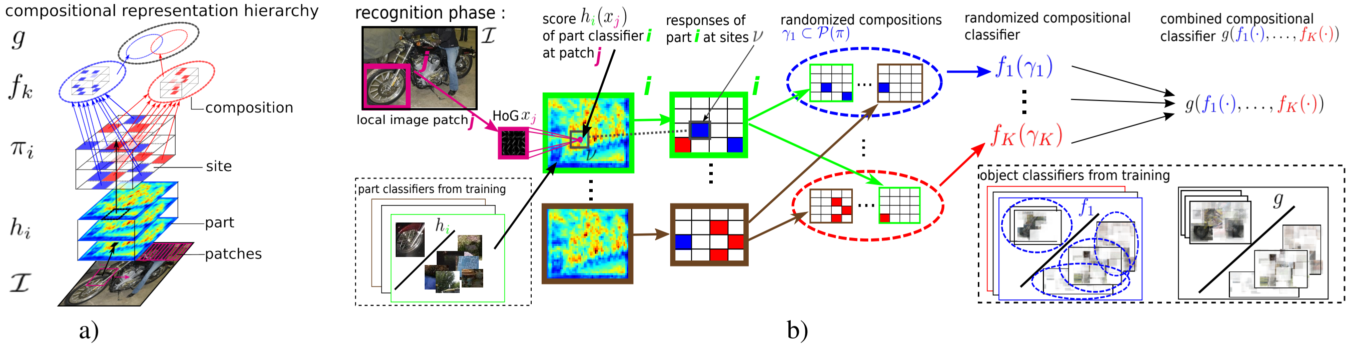
compositional representation hierarchy
g
f_k
π_i
h_i
$\mathcal{I}$
composition
site
part
patches
a)
recognition phase :
$\mathcal{I}$
local image patch j
HoG x_j
score $h_i(x_j)$ of part classifier i at patch j
i
ν
responses of part i at sites ν
randomized compositions $\gamma_1 \subset \mathcal{P}(\pi)$
randomized compositional classifier
$f_1(\gamma_1)$
$f_K(\gamma_K)$
combined compositional classifier $g(f_1(\cdot), \dots, f_K(\cdot))$
$g(f_1(\cdot), \dots, f_K(\cdot))$
part classifiers from training
h_i
object classifiers from training
f_1
g
b)

Figure 2. a) shows the compositional representation hierarchy and b) the detection procedure of our randomized max-margin compositions. Part classifiers responses are pooled at different locations before aggregating them in randomized, discriminatively trained compositions. All compositions then join in a final combined classifier $g(\cdot)$.

proved generalization ability. Compositionality [15, 23] is a powerful principle for reducing the representational complexity to render learning of structured models feasible. We deviate from the common rationale of compositional hierarchies [13, 15, 17, 24, 28] that establish meticulously arranged, semantically meaningful compositions. Rather we show that multiple overlapping *randomized* compositions trained using a max- approach generalize significantly better to new category instances compared to the original parts and thus yield improved performance. Compositions are then all combined by a final non-linear decision function in a third layer of this hierarchy of discriminative classifiers, with part classifiers and the compositional classifiers in the two preceding stages (see Fig. 2).

We thoroughly evaluate the individual contributions and crucial modeling decisions of our model. Experiments are conducted on the well-established, competitive benchmark detection challenges of PASCAL VOC 2007, using the VOC 2010 evaluation server [9], and on the challenging MITIndoor scene recognition dataset of [27]. Our randomized max-margin compositions (RM^{2}C) show, to the best of our knowledge the currently best performance using only HOG features for single class object detection, i.e. without any postprocessing exploiting interactions of multiple object classifiers trained for different classes.

Moreover, the experimental analysis underlines the necessity of large numbers of specific parts because of their mutual unrelatedness and low generalization ability. We also observe that randomly sampling compositions significantly outperforms individual parts, a location based part grouping, and a clustering based on visual similarity. Finally, we show that our approach not only localizes object bounding boxes, but that, although being discriminative, it parses their content to thoroughly explain a test object with the randomized compositional model (cf. Fig. 1). We then propose a novel evaluation setup for part-based models on PASCAL VOC 2010 that allows measuring the accuracy of arbitrary individual parts. We believe that this new experimental protocol is crucial to thoroughly evaluate the intermediate components of hierarchical part-based methods.

2. Related Work

A popular and powerful approach for discriminative part-based object recognition is the deformable part model (DPM) suggested by Felzenszwalb *et al.* [12]. The model trains a latent support vector machine to discover the hidden locations of a fixed number of parts. Zhu *et al.* [32] extended this idea and suggested a deeper hierarchy of parts which is trained using a structural SVM. Recently Song *et al.* [30] suggested a discriminative and-or tree model to automatically learn the configuration of parts. Since the spatial configuration needs to be learned in the training phase, the number of parts is quite restricted. This results in a small set of very general parts that typically correspond to a whole aspect. Contrary to this, our framework is able to handle a very large number of specialized parts. Due to the great number of parts our approach can not only detect object bounding boxes but also provides a parsing of its content (see Fig. 1). Endres *et al.* [8] are avoiding a structured model and use a simple method that pools part responses over proposed object regions with a boosting classifier. Similarly to our approach they start by using part-based exemplar SVM [22]. However, one of the main challenges solved in [8] is how to refine these simple but specialized classifiers to get a smaller more general set of part-classifiers. In addition there has been work on incorporating strong supervision to train part-based object detection models such as [1] and [3] and on different classifiers such as Random Forests [4].

Part-based approaches are recently also becoming more popular for scene classification. Pandey *et al.* [25] adapted the deformable part model for scene classification. On the other hand there are holistic representations such as object bank [20] that require a supervised training of object classifiers. Compositional hierarchies [15, 23] have been proposed to bridge the large gap between local features or parts and the whole object or scene. The fundamental goal is to establish one or more successive representational layers by grouping parts, thus obtaining a hierarchy of successively larger and more meaningful compositions [13, 15, 17, 28]. In contrast to this delicate assembly of compositions, which is common to these approaches, we show that randomized discriminative compositions are ideal for robust aggregation of specialized parts, thus yielding significant performance improvements.

Similar to the discriminative training of intermediate compositions in [24], [29] train mid-level patch classifiers. [16] followed this idea but started from individual exemplar SVM classifiers which are used to mine more positive samples instead of performing an unsupervised clustering as in [29]. Since [29] and [16] are discovering parts in an unsupervised manner they need to solve the problem of finding a good positive training set for parts using clustering, positive mining etc. which is as difficult as the scene classification problem itself. Therefore, our aim is not to make parts more general, but rather to train compositions that generalize better than the specialized part classifiers they aggregate.

3. A Compositional Approach to Discriminative Part-based Recognition

Let us assume for now that we have semi-local features and part classifiers that are specifically trained for individual instances of an object category. We discuss the training of these parts in Sect. 3.3 and provide the classifiers on the project site [1]. Due to the specific nature of such parts, a large number of them is necessary to capture all relevant characteristics of complex object categories. However, training a powerful discriminative model, *e.g.*, a non-linear classifier, on a limited training set, is not feasible based on the high-dimensional combination of a large number of parts. To avoid overfitting we aggregate parts in fewer, overlapping compositions, each capturing a previously learned, random set of parts. These compositions, that can be shared across instances of a category, are all gathering different observations due to the random selection of parts and thus generalize better to novel samples. Sect. 3.1 presents our compositional model before discussing part classifiers and their training in the following sections.

3.1. Randomized Max-Margin Compositions

Assume we have already trained a large set of part classifiers (typically around $P = 1000$ per category), which will be described in Sec. 3.3. For some image site ν the classifier of part i is evaluated densely within this region and the detection scores are pooled yielding a response $\pi_i(\nu) \in \mathbb{R}$ as will be discussed in Sect. 3.2 . At each image site all parts are evaluated. The common approach is then for all sites $\nu \in \mathcal{I}$ on a regular grid within an object bounding box $\mathcal{I}$ to concatenate all part responses. Given the large number of parts this would yield a very high dimensional representation (on average far beyond 20 000-D). In light of the curse of dimensionality, learning object models with this high dimensional representation on a limited set of positive training samples (for PASCAL VOC typically on the order of 100) is inappropriate. One might speculate that there is significant redundancy when a large number of part classifiers is applied to an object, so that grouping related parts or subspace methods could significantly reduce dimensionality. However, since each part classifier represents a single positive object region (Sec. 3.3), we observe that their responses are highly uncorrelated (cf. Fig. 3). Consequently, applying principle component analysis, 90% of the original dimensionality retains only about 40% of the variance.

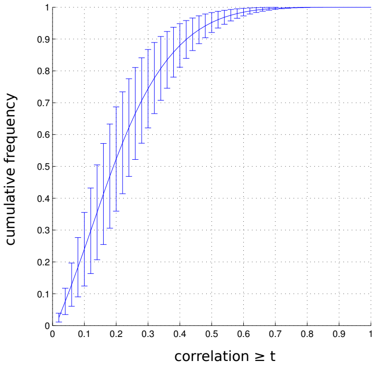

Figure 3. Maximal absolute correlation of a part to any other part evaluated over all categories of VOC 2007. Most parts are highly uncorrelated, i.e. 95% of parts have a correlation of less than .5 to any other.

The $\pi_i(\cdot)$ are essentially trained to act as specialists, each specifically trained for an individual part instance from training. Therefore, we propose to group the responses of all parts i at sites ν to create K groups of part responses $K << P$. Each comprises a large number of part responses and thus generalizes better than individual parts to the large number of instances from an object category. More precisely, let $\pi := \{\pi_i(\nu), \forall i, \nu\}$ and $\mathcal{P}(\pi)$ be the powerset of all responses then we seek K compositions $\gamma_k \subset \mathcal{P}(\pi)$. When applying a composition to a candidate object bounding box $\mathcal{I}$, we obtain a $|\gamma_k|$-dimensional response $\gamma_k(\mathcal{I})$. Following upon the part classifiers, the groups establish a second level in a classifier hierarchy. To render the learning problem feasible, this second level is comprised by linear classifiers $f_k(\cdot)$ trained with hinge loss in a max-margin fashion,

$$\min_{\mathrm{w}} \frac{1}{2}\|\mathrm{w}\|_2^2 + C \sum_{\mathcal{I} \in \mathcal{T}} \max(0, 1 - y_{\mathcal{I}} f_k(\gamma_k(\mathcal{I}))) \qquad (1)$$

where $f_k(\gamma_k(\mathcal{I})) = \mathrm{w}_k^T \gamma_k(\mathcal{I}) + \mathrm{b}_k$, $\mathcal{T}$ denotes the set of training bounding boxes and $y_{\mathcal{I}} \in \{-1, 1\}$ is the class label of the bounding box $\mathcal{I} \in \mathcal{T}$. Now the questions remains,

[1] hci.iwr.uni-heidelberg.de/COMPVIS/research/RM2C

how to obtain the γ_k. From the experiment in Fig. 3 we see that the appearance-based part responses $\pi_i(\nu)$ at locations ν are uncorrelated. Without any extra annotation as in [3] we can from this experiment already suspect that an unsupervised grouping of parts based on their appearance and location will not be desirable. And indeed, combining parts based on similarity in appearance and location using agglomerative clustering (Wards method) does not yield a significant improvement of groups compared to their constituent parts. We experimented with different grouping strategies and measured the performance of the first level compositional classifiers $f_k(\cdot)$ in terms of average precision on a validation set. Fig. 4 shows the cumulative frequency of group classifiers $f_k(\cdot)$, i.e., the fraction of classifiers that succeed a certain average precision. When grouping parts based on their location we observe little gain over the baseline of singleton part groups. An agglomerative clustering based on visual similarity yields a larger improvement over the individual part performance. To achieve a further significant gain we propose to randomize the formation of compositions. Therefore mutually overlapping part response vectors γ_k are drawn randomly from $\mathcal{P}(\pi)$. To simplify their subsequent combination, we demand all γ_k to have a fixed size $|\gamma_k| = L$. Crossvalidation has shown L=3000 (part,location) pairs to yield optimal performance, but the fluctuation within reasonable range was insignificant. Fig. 4 shows that randomized compositions generalize significantly better than clustering parts based on their visual similarity. One might conclude that randomization avoids overfitting by not using visual information twice, i.e., for defining the part classifiers and for clustering them based on visual similarity.

Now we have a manageable number of compositions, each being significantly more informative than the large number of initial parts. Thus, training a non-linear classifier $g(f_1(\cdot), \dots, f_K(\cdot))$ that establishes a third level in the already existing hierarchy of classifiers becomes feasible. Let $F(\cdot) = (f_1(\cdot), ..., f_K(\cdot))^\top$ be the low dimensional feature descriptor that concatenates the K decision values (we use $K = 50$) $f_k(\cdot)$ of the second level group classifiers. The final third level classifier is then trained by optimizing

$$\max_\alpha \sum_{\mathcal{I}\in\mathcal{T}} \alpha_{\mathcal{I}} - \frac{1}{2}\sum_{\mathcal{I}\in\mathcal{T}}\sum_{\mathcal{I}'\in\mathcal{T}} \alpha_{\mathcal{I}}\alpha_{\mathcal{I}'} y_{\mathcal{I}} y_{\mathcal{I}'} \kappa(F(\mathcal{I}), F(\mathcal{I}')) \quad (2)$$

with the radial basis function (RBF) kernel given by $\kappa(F(\mathcal{I}), F(\mathcal{I}')) = \exp(-\frac{\|F(\mathcal{I})-F(\mathcal{I}')\|_2^2}{2\sigma^2})$. The decision function is $g(F(\mathcal{I})) = \sum_{\mathcal{I}'\in\mathcal{T}} \alpha_{\mathcal{I}'} y_{\mathcal{I}'} \kappa(F(\mathcal{I}), F(\mathcal{I}'))$.

3.2. Part Responses on Image Sites

Evaluating a part classifier i only once per image site ν leads to noisy results, since the regular spatial grid of sites is too coarse to deal with local deformations. If a part in an image would be shifted or scaled, so that it is not aligned with a site ν we might miss it. Therefore, we follow common practice and sample local features x_j densely using a sliding window at all locations/scales $j \in \nu$ within sites. To get the sites we use regular grids of size 1×1, 2×2 and 4×4. As feature we use HOG and for the j we use the location/scale pyramid of [12]. As a result we obtain classifier scores $h_i(x_j)$ for each part (cf. Sect. 3.3). The part response to a site is then defined by max pooling over all locations/scales within ν,

$$\pi_i(\nu) = \max_{j\in\nu} h_i(x_j). \quad (3)$$

This aggregation of part responses on a spatial grid has been shown to work well in different vision problems [19, 20, 29].

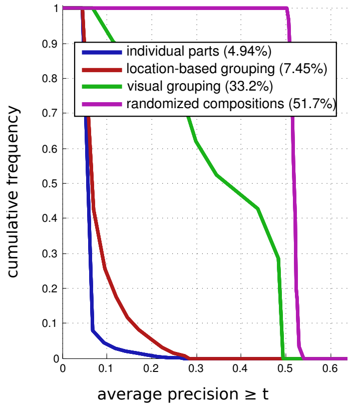

Figure 4. Comparing different grouping strategies for assembling compositions on VOC 2007 *bicycle*. Cumulative frequency of group classifiers $f_k(\cdot)$ w.r.t. their average precision.

3.3. Learning Parts without Part Annotation

Learning part models without annotation of parts is a challenging problem. Without extra annotation, the task of finding corresponding parts in different object bounding boxes turns out to be as difficult as finding the object itself, since the locality of parts leads to ambiguities. Thus parts are typically detected conjointly, linked by a spatial model that enforces spatial consistency. However, when learning a part, we have neither an object model provided nor any other parts. Thus, finding all instances of a part in all training images is daunting. And indeed it was shown that clustering based on the distance of features (*e.g.* HOG) is not very reliable [14, 29]. The problem is then that incorrect groups of parts at this initial stage will lead to mistakes that accumulate during later stages. We therefore train part models with just a single positive sample and a set of negatives as suggested by [22]. To obtain the positive part samples we randomly select a large number of patches at different locations and scales within training bounding boxes. All parts together should exhibit a good coverage of all training images. Therefore, we do not want to get very similar patches with high overlap in the same bounding box and therefore restrict the overlap between sampled patches in the bounding box to be less than 20%. Additionally we restrict the number of parts per box and sample a maximum of 20 parts. Note, that significantly less parts maybe sampled if the object bounding box is very small. Now we have one positive sample x_p per part, and similar to [22] we perform negative mining on up to 2500 images to obtain a set of negatives $\mathcal{N}$. The corresponding classification function h_i is

$$\min_{\omega} \frac{1}{2}\|\omega_i\|_2^2 + C_1 \max(0, 1-h_i(x_p)) + C_2 \sum_{x\in\mathcal{N}} \max(0, 1+h_i(x)) \tag{4}$$

were $h_i(x) = \omega_i^T x + \beta_i$. The part features x are HOG descriptors [7] using 25 cells that are fitted to the part as in ESVM [22]. The number of pixels per cell depends on the scale on which the part was sampled. The minimum cell size is 4 pixels. In our framework the trained exemplar SVMs act as specialized parts. One might think that a part classifier trained on one positive sample is overfitting badly and therefore performance of the individual parts might be very poor compared to more general parts using a larger set of positive training samples. To get an idea of the quality we are evaluating the individual performance of the part classifiers in the next section.

Recognition Phase To perform object detection in a novel test image (see Fig. 2b) we first need to extract HOG descriptors x_j and run part classifiers $h_i(x_j)$. Then we pool part responses using Eq. 3 into $\pi_i(\nu)$ before running the composition classifiers $f_k(\cdot)$. Finally we evaluate $g(F(\cdot))$ to combine all compositions using the non-linear classifier.

3.4. Part Evaluation

To evaluate the performance of our part classifiers we are using the keypoint annotation of [2] for the PASCAL 2010 dataset. However, in contrast to poselets this is here merely for our subsequent evaluation and not for training. Since our parts are trained in an unsupervised manner using HOG features we are comparing the performance of our parts to those of the Deformable Part Model (DPM) [12] which are using a similar setup. In contrast to our parts the DPM parts are much more general since they are trained on all training images from an aspect of a category.

To evaluate the detection performance of individual parts we first need to generate groundtruth on which we can test. In contrast to [3] there are no annotations specific to our parts, but the idea is to measure how much a part shifts between training and testing relative to the existing keypoint annotation of [3]. For the positive training sample x_p that defines the part we therefore measure its euclidean distances to all keypoints within the object bounding box. During detection we again compute the distances to the same keypoints. Comparing the training and test vector of keypoint distances thus defines a similarity measure. Now we can rank parts according to their mean average precision, i.e., how good they are in detecting a similar object region as they were trained upon, where similarity is measured with respect to annotated semantic landmarks from [3]. Fig. 5 compares the 48 DPM parts [12] with the randomly sampled parts from Sect. 3.3 (also a subset of 48). We observe that in the large pool of weak parts there is still a sufficient number of parts that have favorable detection performance compared to the DPM parts.

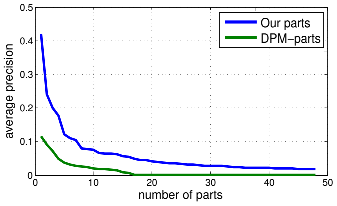

Figure 5. Performance comparison of the 48 DPM parts with our randomly sampled parts (also a subset of 48) in terms of average precision, see Sect.3.4

4. Experiments

In our experiments we are providing object recognition and scene classification results on three of the most challenging datasets. For object recognition we are evaluating our approach on PASCAL VOC 2007 and 2010 . The scene classification results are evaluated on the MITIndoor dataset [27]. Our experimental results show competitive performance to recent state-of-the-art part based approaches on all datasets.

4.1. Object Recognition

We follow the standard training and testing protocols for the PASCAL detection challenge only using provided bounding box annotation on the object category level. Additionally we are showing qualitative results in terms of a back-rendering of our training parts in the detection box to visualize how our model is explaining objects (see Fig. 6).

4.1.1 Implementation Details

Training Since we are training classifiers on a part level and on an object level we need to split the training data, to avoid over-fitting. Considering our part classifiers are trained in an exemplar fashion over-fitting is not an issue on the rare positive samples as one part classifiers is only over-fitting in one image at a certain location and scale. The sampling of positive patches described in Sec. 3.3 can therefore be performed on the whole trainval set. Since each part classifier is performing a negative mining, the part classifiers might over-fit when we are applying them on the same negative images again to get the response maps. Therefore we are only using 2500 negative images from the PASCAL training data for the negative mining procedure of the part classifiers. To train the object classifiers (i.e. $f_k(\cdot)$ and $g(\cdot)$) we use all the positives form the trainval set and all negative images remaining after training the part classifiers. To

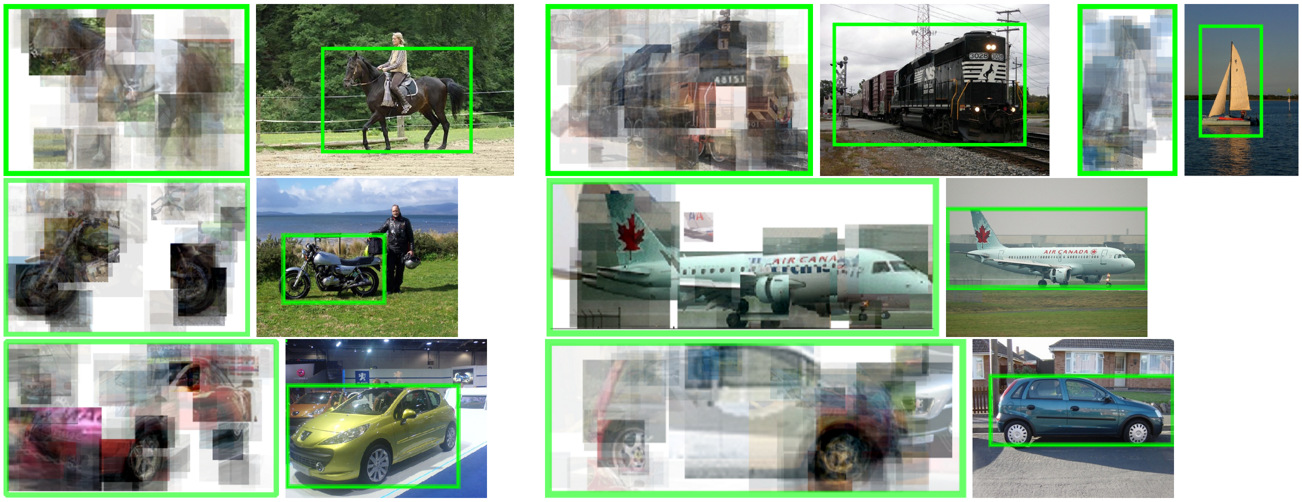

Figure 6. Example reconstructions of true positives using compositions and parts from positive training samples.

get a set of hard negative samples we apply the deformable part model with a low threshold (-1.1) and use the resulting false detections. Note, that we use the same models and parameters for hypothesis generation at detection time. For training the SVM classifiers we use LIBSVM to train nonlinear classifiers and otherwise LIBLINEAR [10].

Part Selection Since we are sampling an over complete set of parts the number of parts can be extremely large for classes with a lot of objects like the person category. This raises the question if all of these parts are actually needed. Therefore we perform an experiment where we use an increasing number of parts (in steps of 100 parts) for training and evaluate the performance on the validation set. Note, that since we evaluate on the validation set only the training data are used to train our framework. We order the parts according to their strength based on the absolute weights of a linear SVM classifier trained on the maximum response of each part per training sample. For each of our evaluations we are using the best N parts for training. Fig. 7 shows that the mean average precision is saturating around 1000 parts. This confirms that a large number of part classifiers is actually needed. One could think that the reason this high number of parts are needed is because the individual performance of our exemplar-based parts is very weak. However, as we were discussing in detail in Sec. 3.4 and is shown in Fig. 5 a subset of our part classifiers is even performing better than the DPM parts. Based on these results we are selecting the subset of parts for each category with the highest performance on the validation set. Detecting with all these classifiers may seem very time consuming. However, the filter operation is just a single dot-product for all the part classifiers. Creating the response maps for 1000 part classifiers takes around 13 seconds. For comparison the DPM [12] takes 7 seconds to create response maps for 54 object and part classifiers. The reason for the comparably small overhead of our system is that the time needed to build HOG features and extract detection windows for an image is significantly higher than the detection time. Therefore, the more filters are used the more favorable it is to first extract HOG features for all windows and perform a single matrix multiplication than performing a separate convolution for each filter as done by the DPM.

4.1.2 Comparison with other Methods

Since we suggest a part-based approach the focus of our evaluation is to compare with other part-based approaches. There exist several methods such as [5, 31] that focus on how the responses of several classifiers can be used to improve overall detection performance. These methods can be applied in a post-processing step for any part based method. Therefore part-based methods are evaluated without context in common literature.

PASCAL VOC2007 Our final approach (RM^{2}C) is also incorporating parts that are root filters. Our results show that the suggested approach already gives state-of-the art performance without applying larger parts corresponding to objects (RM^{2}C w/o obj.). Additionally we are comparing our approach to three other part-based approaches. All approaches are utilizing HOG features as a low level representation. The detection results are summarized in Tab. 1. Our method outperforms all other approaches on 17 out of 20 categories. Significant improvements are reached on articulated objects as dogs (8.1%), cats (6.5%) and birds (2.5%). However, also more rigid objects with high intra class variability benefit from our specialized part-classifier compositions as aeroplanes (4.5%) and tvmonitors (2.5%). In mean we are gaining 1.9% over the And-Or Tree (AOT), 2.9%

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
DPM rel5 [12]	33.2	60.3	10.2	16.1	27.3	54.3	**58.2**	23.0	20.0	24.1	26.7	12.7	58.1	48.2	**43.2**	12.0	21.1	36.1	46.0	43.5	33.7
LHS [32]	29.4	55.8	9.4	14.3	28.6	44.0	51.3	21.3	20.0	19.3	25.2	12.5	50.4	38.4	36.6	15.1	19.7	25.1	36.8	39.3	29.6
AOT [30]	35.3	60.2	9.4	16.6	29.5	53.0	57.1	23.0	22.9	27.7	28.6	13.1	58.9	49.9	41.4	**16.0**	22.4	37.2	48.5	42.4	34.7
RM^{2}C w/o obj	37.0	58.3	12.0	14.7	22.9	51.3	51.7	23.7	21.7	25.0	29.0	20.6	51.4	46.1	36.3	12.7	22.3	35.1	43.9	41.8	32.9
RM^{2}C	**37.7**	**61.4**	**12.7**	**17.6**	**29.9**	**55.1**	56.3	**29.5**	**24.6**	**28.2**	**30.7**	**21.2**	**59.5**	**51.5**	40.3	14.3	**23.9**	**41.6**	**49.2**	**46.0**	**36.6**

Table 1. Performance comparison using average precision (AP) for the PASCAL VOC2007 dataset. For abbreviations see Sect. 4.1.2

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
DPM rel5 [12]	45.6	49.0	11.0	11.6	27.2	50.5	43.1	23.6	17.2	23.2	10.7	20.5	42.5	44.5	41.3	8.7	29.0	18.7	40.0	34.5	29.6
Poselets [3]	33.2	**51.0**	8.5	8.2	**34.8**	39.0	**48.8**	22.2	-	20.6	-	18.5	**48.2**	44.1	**48.5**	9.1	28.0	13.0	22.5	33.0	-
BCP [8]	44.3	35.2	9.7	10.1	15.1	44.6	32.0	**35.3**	4.4	17.5	**15.0**	**27.6**	36.2	42.1	30.0	5.0	13.7	18.8	34.4	28.6	25.0
AOT [30]	44.6	48.5	10.8	12.9	26.3	47.5	41.6	21.6	**17.3**	23.6	11.5	22.9	40.9	45.3	37.9	9.6	**30.4**	25.3	39.0	31.2	29.4
RM^{2}C	**49.8**	50.6	**15.1**	**15.5**	28.5	**51.1**	42.2	30.5	**17.3**	**28.3**	12.4	26.0	45.6	**51.8**	41.4	**12.6**	**30.4**	**26.1**	**44.0**	**37.6**	**32.8**

Table 2. Performance comparison using average precision (AP) for the PASCAL VOC2010 dataset. Note that our approach outperforms Poselets comparing the mean of the 18 classes where the detection results are provided by 5.3%.

over the Deformable Part Model (DPM) and 7% over the Latent Hierarchical Structures (LHS).

Since the number of random compositions (K=50) is rather small one could suspect that the variance of the detection performance is high. However, measuring the variance of the mean average precision of five different random composition samplings showed a favorable variance of about 0.1%.

PASCAL VOC2010 Additionally, we are providing results on the PASCAL VOC2010 dataset were we outperform other approaches on 12 out of 20 classes (see Tab. 2). Our approach performs particularly well for classes that can be considered as very difficult due to the huge intra-class variations as birds, boats and potted plants were the improvement is up to 4.1% in terms of average precision. The comparison with the Boosted Collection of Parts (BCP) is particularly interesting, since due to their usage of exemplar parts it is the most similar approach to our compositional part-model. We are showing superior performance on 17 out of 20 classes, improving the average precision by 7.8%. While poselets are giving the best performance on 5 out of 20 classes they also perform more than 10% worse than our detection system on 5 other classes. Note that we are outperforming the poselets even though this approach uses additional ground truth annotation in the form of keypoints for training while ours only depends on bounding box annotations at object level. Comparing the mean over the 18 classes on which results for the poselets are available we outperform them by 5.3%. All in all we are gaining 3.2% in terms of mean average precision over the DPM which is the best performing approach we are comparing to.

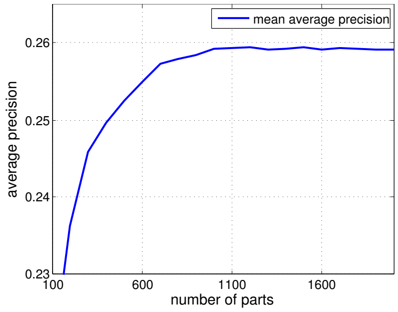

Figure 7. Mean average precision of all classes of the PASCAL VOC2010 dataset on the validation set training our model with different number of parts.

4.2. Scene Classification

For scene classification we are using the protocol given in [27] where each scene class consists of 80 training images and 20 test images. We provide results as in [27] in terms of classification accuracy obtained by averaging the diagonal of the confusion matrix and in terms of mean average precision which is used as an additional measurement in [16].

We compare our performance to 7 different classification approaches (see Tab. 3). The focus of our evaluation is the comparison to other methods that are using semantical part classifiers based on HOG features for scene classification. Therefore the most important comparisons are in the lower half of Tab. 3, since these approaches are methodologically most similar to the one we are suggesting. Our results show that we outperform Mid-Level Patches [29] by 13% and the Bag of Parts (BoP) by 5% in term of classification accuracy. The improved fisher vectors (IFV) can be combined with all part based approaches to boost performance as it was done by IFV+BoP [16]. Since we are outperforming the individual performance of BoP, it should be expected that

Method	Acc. (%)	Mean AP
Object Bank [20]	37.60	-
RBoW [26]	37.93	-
DPM+GIST-color+SP [25]	43.10	-
Patches+GIST+SP+DPM [25]	49.40	-
IFV+BoP [16]	63.10	63.18
Mid-Level Patches [29]	38.10	-
BoP [16]	46.10	43.55
RM^{2}C	51.34	46.70

Table 3. Average classification performance on the MITIndoor Dataset. Upper half of the table shows diverse approaches for scene classification while the lower half focuses on approaches using semantic parts and are therefore most similar to our approach.

the combination with fisher vectors would outperform their combined approach. However, the aim of this experiment was to compare our method with other related part based approaches.

5. Conclusion

We have proposed a compositional approach that can integrate large numbers of weak parts in a strong discriminative model. Contrary to the main theme of the filed, we randomly sample instance specific parts and randomly aggregated them in compositions that are trained using a max-margin procedure. The approach has shown favorable performance on standard benchmark datasets for object detection and scene classification and the potential of its constituents has been evaluated individually.[2]

References

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

[2] L. Bourdev, S. Maji, and J. Malik. Detection, attribute classification and action recognition of people using poselets (in submission). In *PAMI*.

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

[4] L. Breiman. Random forests. *Machine Learning*, 2001.

[5] G. Chen, Y. Ding, J. Xiao, and T. Han. Detection evolution with mulit-order conextual co-occurrence. *CVPR*, 2013.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *ECCV Int'l Workshop on Statistical Learning in Computer Vision*, 2004.

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

[8] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. *CVPR*, 2013.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[10] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 35(8), 2013.

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[13] S. Fidler, M. Boben, and A. Leonardis. A coarse-to-fine taxonomy of constellations for fast multi-class object detection. *ECCV*, 2010.

[14] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. *ECCV*, 2012.

[15] J. Jin and S. Geman. Context and hierarchy in a probabilistic image model. *CVPR*, 2006.

[16] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Ziserman. Blocks that shout: Distinctive parts for scene classification. *CVPR*, 2013.

[17] I. Kokkinos and A. Yuille. Inference and learning with hierarchical shape models. *IJCV*, 93:201–255, 2010.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS'12*.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006.

[20] L.-J. Li, H. Su, e. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. *NIPS*, 2010.

[21] S. Maji and J. Malik. Object detection using a max-margin hough transform. *CVPR*, 2009.

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

[23] B. Ommer and J. M. Buhmann. Learning compositional categorization models. *ECCV*, 2006.

[24] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual object categories for recognition. *PAMI*, 32(3):501–516, 2010.

[25] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. *ICCV*, 2011.

[26] S. N. Parizi, J. Oberlin, and P. F. Felzenswalb. Reconfigurable models for scene recognition. *CVPR*, 2012.

[27] A. Quattoni and A. Torralba. Recognizing indoor scenes. *CVPR*, 2009.

[28] M. Ranzato, V. Mnih, J. M. Susskind, and G. E. Hinton. Modeling natural images using gated MRFs. *PAMI*, 35(9):2206–2222, 2013.

[29] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. *ECCV*, 2012.

[30] X. Song, t. Wu, Y. Jia, and S. Zhu. Disriminatively trained and-or tree models for object detection. *CVPR*, 2013.

[31] Z. Song, Q. Chen, Z. Huang, A. Hua, and S. Yan. Contextualizing object detection and classification. *CVPR*, 2010.

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

[2] This research has been funded in part by the Ministry for Science, Baden-Wuerttemberg.