Adding discriminative power to hierarchical compositional models for object class detection

Matej Kristan[1], Marko Boben[1], Domen Tabernik[1], and Ales Leonardis[2 1]

[1] Faculty of computer and information science,University of Ljubljana,
matej.kristan@fri.uni-lj.si,
WWW home page: http://vicos.fri.uni-lj.si/matejk/
[2] CN-CR Centre, School of Computer Science, University of Birmingham

Abstract. In recent years, hierarchical compositional models have been shown to possess many appealing properties for the object class detection such as coping with potentially large number of object categories. The reason is that they encode categories by hierarchical vocabularies of parts which are shared among the categories. On the downside, the sharing and purely reconstructive nature causes problems when categorizing visually-similar categories and separating them from the background. In this paper we propose a novel approach that preserves the appealing properties of the generative hierarchical models, while at the same time improves their discrimination properties. We achieve this by introducing a network of discriminative nodes on top of the existing generative hierarchy. The discriminative nodes are sparse linear combinations of activated generative parts. We show in the experiments that the discriminative nodes consistently improve a state-of-the-art hierarchical compositional model. Results show that our approach considers only a fraction of all nodes in the vocabulary (less than 10%) which also makes the system computationally efficient.

Keywords: compositional models, hierarchical models, categorization, discriminative parts

1 Introduction

Object detection and categorization is a challenging problem in computer vision. Flat constallation models, in which object features are directly aggregated to form an object, have been shown to perform well on object class detection [1–3]. With the goal of extending detectors to many categories, scalability, storage and efficient inference issues have become an important aspect of the detection process. A family of models have recently emerged to address these issues, most notably in a form of multilayered hierarchies such as recursive compositional hierarchies [4–7] AND-OR graphs [8], sum-product networks [9] and convolutional networks [10, 11].

A central point of the hierarchical models is that their lowest layer is composed of elementary parts, which are combined to produce more complex parts

on the next layer. This procedure may be recursively repeated over several layers which gradually increases the complexity of the vocabulary of parts. An appealing aspect of the compositional hierarchies is that, on the one hand, they offer sharing of object parts within object category, while on the other hand they can also reuse the parts at multiple levels of granularity among different categories [12]. In fact, in the recent work Fidler et al. [12] have shown that a hierarchical compositional model allows incremental training and significant sharing of parts among many categories. The sharing reduces the storage requirements and at the same time makes inference efficient, since hypotheses of the shared parts are verified simultaneously for multiple categories. Note also that [12] performs detection without resorting to a sliding-window approach.

Hierarchies such as [4, 6, 7] are in their nature generative in that they optimize their structure to sparsely reconstruct the observed objects. In presence of clutter, however, this results in many spurious detections. This also presents a difficulty when trying to discriminate between visually similar categories. For example, on an image of a cow, a generative hierarchy can robustly detect a cow, but might also with similar certainty detect a horse at the same location. The latter qualifies as spurious detection as illustrated in Figure 1. From a perspective of a hierarchy, a cow and horse category may share multiple parts and differ only in a small subset of all parts required for their detection. Such behavior is inherent to many generative hierarchical approaches to category detection. In [13] a related problem has been addressed in a patch-based constellation model. There the authors showed that detecting differences between visually similar categories requires first detecting salient features which may not be discriminative on their own, but can be detected with a high certainty. Using the salient features as anchor points, more illusive features, called the satellite features, were detected and these both types of features together resulted in improved discrimination. An advantage of a generative hierarchical model [12] over a part-based detector [13] is that it already encodes the object's visual appearance as vocabulary of parts at various levels of granularity. Some of these parts can be identified as the subtle differences between the object categories, but this is not exploited in the existing standard generative hierarchy [4]. This paper presents a novel method that seeks to combine the best properties from reconstructive hierarchical part-based models and the purely discriminative models. Our scientific hypothesis is that the discrimination performance of a generative hierarchical compositional model can be improved by identifying and using the discriminative parts of the existing hierarchy's vocabulary for a detection hypothesis verification.

1.1 Our approach

As our main contribution, we introduce an additional discriminative analysis over an existing reconstructive hierarchical model to improve its discriminative properties. For each pair of categories modeled in the generative hierarchy, we form a so-called *discriminative node*. A discriminative node is a weighted linear combination of cumulative responses of the *reconstructive nodes* (parts) that are activated at detection of a category. To prevent overfitting, we seek a sparse

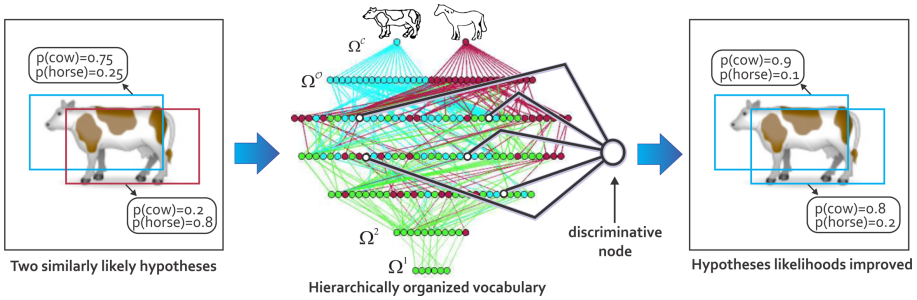

Fig. 1. The left image depicts two competing hypotheses at a location of a cow – one correct detection and one spurious detection. The middle image illustrates our contribution – a discriminative node connects several nodes from hierarchically organized reconstructive vocabulary and forms a discriminant verification of the hypotheses. The right image shows the rescored hypotheses which result in improved detection/categorization.

solution to the weight vectors that form the discriminative nodes. We therefore cast the problem of determining the weights as a sparse logistic regression problem [14]. At detection stage, the hierarchy predicts a bounding box for an object. The responses of the activated parts within the predicted bounding box are used by the discriminative node to verify, or rescore, the detection. An example of the discriminative node, along with its connections to vocabulary of parts, is illustrated in Figure 1. There are several appealing properties of the proposed approach over the discriminatively trained hierarchies for object detection. Our approach does not require re-training the entire vocabulary from scratch when a new category is introduced and does not hamper the incremental training property [12] of the generative hierarchical model. This is because our network of discriminative nodes does not work on pixel level, but rather on the existing vocabulary parts in the reconstructive hierarchy. Therefore, as a new category is introduced, and new parts are formed in the hierarchy, we only require forming new discriminative nodes between the existing categories and the newly added category. This is an advantage over the feed-back methods that adapt parts in discriminative training: Assume that we have learned the parts for cows, horses, sheep etc., and want to discriminate between cows and horses. Since the parts are shared among many categories, adapting them to only discriminate between cows/horses (feed-back approach) will hinder detection of other categories that share these parts. Our approach rather reinterprets the importance of existing parts for a specific discrimination. Since the discriminative node is simply a sparse linear combination over the node responses that are activated during detection process it maintains the efficient inference properties studied in [12]. We apply the concept of discriminative nodes on a state-of-the-art hierarchy called the *learned Hierarchy of Parts* (lHoP) [4]. Experimental results show that the discriminative nodes significantly reduce spurious detections in the lHoP. We

demonstrate this on examples of discriminating between visually similar categories as well as discriminating object categories from the background.

The remainder of the paper is structured as follows. Section 2 overviews the lHoP model [4] and Section 3 introduces the concept of discriminative nodes. In Section 4 we experimentally show consistent improvement of lHoP by our discriminative nodes and draw conclusions in Section 5.

2 The learned hierarchy of parts – lHoP

The learned hierarchy of parts (lHoP) [4] is a recursive compositional vocabulary of shape parts represented by a directed graph (Figure 2, left). The top layer of the hierarchy is called the categorical layer Ω_C and contains a single categorical node per category. Each categorical node is represented by a distribution over various shapes of the corresponding category. In practice this means that each categorical node is connected by an *or* connection to several nodes at one layer lower in the hierarchy – the object layer Ω_O. This makes a categorical node a mixture model over alternative object hierarchies. A hierarchy associated to an object node is recursively composed: the object node is a composition of several child nodes located at one layer lower in the hierarchy, the recursive composition rule applies to each of its child nodes and follows down to the lowest layer Ω_1. In this respect, all layers together form a hierarchically encoded vocabulary $\Omega = \Omega_1 \bigcup \cdots \bigcup \Omega_C$. The entire vocabulary Ω, along with the vocabulary parameters is learnt from the training set of images. In a standard setting, the vocabulary of the layer 1 is fixed and is composed of Gabor filters positioned at six orientations. By virtue of composition, vocabulary of each layer is composed of combinations of parts from the previous layers. The next two layers are learnt jointly for all categories in an unsupervised fashion from the training images . The remaining layers are then trained sequentially, one category at a time. This has the effect that the vocabulary parts up to layer 3 are densely shared among the categories. While the higher layers become more specialized, the parts are still largely shared among the different categories [4].

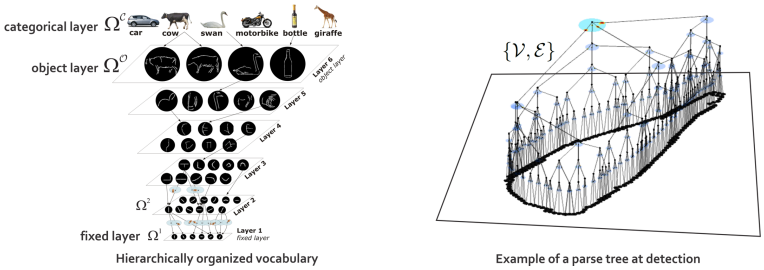

Fig. 2. Illustration of a hierarchically organized vocabulary of parts in the lHoP (left) and an example of the detected parse tree along with the detected vocabulary parts at various locations (right). (Images reprinted with permission from [4])

At detection stage, a parse tree $\{\mathcal{V}, \mathcal{E}\}$ (Figure 2, right) is formed over an image region $\mathbf{I}$, with vertices $\mathcal{V}$ representing the object parts μ and edges $\mathcal{E}$ representing the parent-child relations between parts (the compositions). The relation $\nu \in ch(\mu)$ means that ν is a child of node μ and $\mathcal{V}_{\text{leaf}}$ denotes a set of leaf nodes in the tree. A state $\omega_\mu = \{\mathbf{x}_\mu, \theta_\mu, P_\mu\}$ of the node μ is defined by its identity from the vocabulary, P_μ, the position $\mathbf{x}_\mu$ and parameters θ_μ that constitute the relative positions of its child nodes. Let μ^O denote an object node in the parse tree. The scoring function over the state W of the tree, conditioned on μ^O and the image $\mathbf{I}$, can be written as

$$p(W|\mathbf{I}, \mu^O) \propto \prod_{\mu \in \mathcal{V}/\mathcal{V}_{\text{leaf}}} p(\omega_\mu | ch(\omega_\mu)) \prod_{\nu \in \mathcal{V}_{\text{leaf}}} p(\omega_\nu | \mathbf{I}), \tag{1}$$

where $p(\omega_\mu|ch(\omega_\mu))$ is a score modeling the spatial deformation of child nodes $ch(\omega_\mu)$ relative to their parent μ and $p(\omega_\mu|\mathbf{I})$ is the likelihood of the leaf node μ given the image measurements. In [4] the spatial scores are defined as Gaussian mixture models, and the likelihoods are defined as responses to the Gabor filter bank. During the inference, parts of the lowest layer (layer 1) are activated first. With some probability, these activations index into a subset of parts in the next layer, which are activated if their conditional probability is above a learnt threshold. In this manner, the activations continue throughout the hierarchy up to the categorical layer. After a nonmaxima suppression operation, a category node is activated at a certain location in the image. By tracing the activated nodes back through the hierarchy to the image level, we obtain the predicted bounding box of the detected category instance.

Since the hierarchy is reconstructive in nature, it may also fire on the image structures that belong to the background. As a result, some activations and their bounding boxes do not necessarily correspond to an actual object instance but are rather spurious detections. For each bounding box, we can obtain its score, or likelihood, from (1). However, this score implicitly measures the coverage of the detection by the vocabulary parts and is as such not necessarily optimal for deciding whether a particular bounding box actually corresponds to one category or the other, or the background. In the following section we propose an approach to identify the parts in the vocabulary, which can be utilized to improve verifications of the predicted bounding boxes.

3 Adding discriminative nodes to a generative hierarchy

A detection from lHoP provides a predicted bounding box of an object category along with its parse tree of node activations. Note that a single part from a vocabulary can activate at multiple locations in the image, thus resulting in multiple activated nodes. For each predicted bounding box we form a cumulative response vector, a histogram over the vocabulary $\mathbf{h} = [h_1, \ldots, h_k, \ldots h_K]$, where the k-th bin in the histogram is a summation of scores of all activated nodes with part index k from vocabulary Ω that contributed to the detected bounding

box,

$$h_k = \sum_{\mu \in \mathcal{V}} \delta_k(P_\mu) p(\mu | ch(\mu)), \tag{2}$$

where $\delta_k(\cdot)$ is the Kronecker delta and $p(\mu|ch(\mu))$ is the score of the detected part at node μ and $\mathcal{V}$ are the vertices of the parse tree. By analyzing $\mathbf{h}$ we can predict which parts of the vocabulary contribute the most to discriminating between instances of different categories. For a pair of categories, i and j, we define a *discriminative node* as a weighted summation over the features of $\mathbf{h}$,

$$f(\mathbf{h}; \Theta^{(i,j)}) = \sum_{k=1}^{K} \theta_k^{(i,j)} h_k + \theta_0^{(i,j)}, \tag{3}$$

where $\Theta^{(i,j)} = [\theta_0^{(i,j)}, \ldots, \theta_K^{(i,j)}]^T$ is the vector of weights defining a linear hyperplane that separates the two categories. The probability of category j, given the observation $\mathbf{h}$, can be written as a logistic function

$$p(j|\mathbf{h}, \Theta^{(i,j)}) = \frac{1}{1 + \exp(-f(\mathbf{h}; \Theta^{(i,j)}))} \tag{4}$$

and so $p(i|\mathbf{h}) = 1 - p(j|\mathbf{h})$. Training the discriminative nodes entails estimation of the weight parameters $\Theta^{(i,j)}$ in (4). Recent research in sparse coding [15] has shown compelling evidence of improved performance of sparse solutions in discriminative training. To find a sparse solution that maximizes the discrimination between pairs of categories, we use the automatic relevance determination approach from [14]. Briefly, a Gaussian prior is placed over each parameter $\theta_k^{(i,j)}$ with zero mean and variance $\alpha_k^{(i,j)}$, i.e., $p(\theta_k^{(i,j)}|\alpha_k^{(i,j)}) = \mathcal{N}(0, 1/\alpha_k^{(i,j)})$, and a noninformative prior is placed over the hyperparameter, i.e., $p(\alpha_k^{(i,j)}) = \alpha^{-1}$. We use the variational approach of [14] to obtain the sparse vector $\Theta^{(i,j)}$.

For the multiclass setting we use a standard one-versus-one scheme in which the discriminative nodes between all pairs of categories are calculated. For the estimation of their weight vectors, we apply the above sparse formulation. At the prediction stage, the final probability of a category c is calculated by multiplying probabilities from all pairwise logistic regressors (4). The probability that the cumulative response vector $\mathbf{h}$ was generated under the category c is therefore

$$p(c|\mathbf{h}, \Theta) = \prod_{(i,j)} p(j|\mathbf{h}, \Theta^{(i,j)})^{\delta_c(j)} [1 - p(j|\mathbf{h}, \Theta^{(i,j)})]^{\delta_c(i)}, \tag{5}$$

where Θ is the set of all pairwise discriminants and the product runs over all category pairs (i, j). A region from which the feature vector $\mathbf{h}$ is extracted is classified as the category $\tilde{c}$ that maximizes the probability

$$\tilde{c} = \arg\max_c p(c|\mathbf{h}, \Theta). \tag{6}$$

Figure 3 illustrates discriminative nodes for discriminating between three categories.

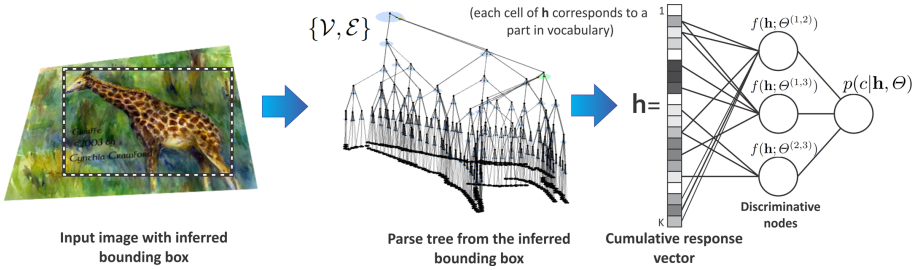

Fig. 3. Cumulative response vector $\mathbf{h}$ is calculated from the parse tree $\{\mathcal{V}, \mathcal{E}\}$. Three discriminative nodes $f(\mathbf{h}; \Theta^{(i,j)})$ are sparsely connected to different parts of $\mathbf{h}$, and discriminate between different category pairs. The nodes are connected to form a probability distribution $p(c|\mathbf{h}, \Theta)$ over the three categories.

4 Experiments

We have performed two sets of experiments to analize the proposed application of discriminative nodes to a generative hierarchy in the context of object detection and categorization. As a baseline hierarchy we have used the lHoP [4]. The first experiment was designed to show improved discrimination between visually-similar categories when using the discriminative nodes with the hierarchy. In the second experiment we demonstrate that the discriminative nodes improve the rescoring stage of lHoP's hypotheses forming for the object category detection. In all experiments, the image edge features were enhanced by preprocessing with the Berkley edge detector [16]. Note that the estimation of the discriminative nodes (Section 3) requires specifying only a single hyperparameter for sparsity of their weight vectors. In all experiments this parameter was automatically selected through cross-validation on the training set. In the following we use dlHoP to refer to the lHoP which uses our discriminative nodes.

4.1 Differences between cows, horses and the "background"

The proposed discriminative nodes can treat the background category exactly the same as any other, tangible, category. In this experiment we therefore demonstrate the performance of the discriminative nodes when discriminating between visually-similar, real-life, articulated categories as well as the background clutter. To this end we have compiled a Cow-Horse dataset by combining the Leeds Cows dataset [17] and Weizman Horses dataset [18]. The Weizman Horse dataset is already split into train and test images. The Cow-Horse train images were therefore composed from the train images from Weizman Horse dataset and a quarter of randomly selected images from the Leeds Cow dataset. The remaining images were used for testing. In this way we obtained a dataset with 75 images for training and 263 for testing. The task of this experiment was to differentiate among two visually-similar categories, and the background, and localize them in the image. Among the detected bounding boxes that overlapped with a ground-truth

bounding box by at least 30% under Pascal criterion, the one with the highest score was taken as a candidate detection. If its predicted class corresponded to the ground truth label, then it was marked as true-positive and false-positive otherwise. All the remaining detections ware marked as false-positives. The lHoP was trained under a weak supervision from the ground truth bounding boxes. It produced a seven-layer hierarchy with, on average, 6, 33, 161, 180, 93, 104 and 2 vocabulary nodes at the corresponding layers. The examples of cows and horses for dlHoP were obtained from the lHoP predicted bounding boxes on the training images that intersected with the ground truth by more than 30%, while the remaining detections were taken as background examples. The lHoP as well as dlHoP processed the test images over four scales at detection stage followed by a nonmaxima suppression.

On average, the dlHoP selected 18, 15 and 11 parts to form discriminative nodes for cow/horse, cow/background, and horse/background, respectively, and approximately eight percent of all distinctive reconstructive parts in vocabulary were selected. The classification results for the lHoP and dlHoP are shown in Table 1. On average, dlHoP outperformed the lHoP by reducing the number of false detections on the background and at the same time improving discrimination between the two categories. In particular, it significantly reduced confusion of cows for horses. The improvement of dlHoP's increased precision at lower false positive rate is evident from the AP score. Figure 4 shows an example of the selected vocabulary parts by the cow/horse discriminative node and the average distribution of the chosen parts across entire hierarchy by all three discriminative nodes. In red and green we show typical and atypical nodes that are selected per category. Presence of typical nodes and atypical nodes increases and decreases the category likelihood, respectively. These nodes were determined by analyzing the weights in the regression vectors. The majority of the selected nodes for discrimination between cow and horse category and the background were selected from layers three to six. By inspecting the most frequently selected parts we can see that the nodes used to discriminate the cow and horse category varied in granularity with some corresponding to partial or full compositions of the cow and horse. Among the nodes for discrimination between cows and the background, we have found compositions of a cow and again parts with lower granularity. Similar holds for discrimination of horses from the background.

Table 1. Confusion matrix for the Cow-Horse dataset for the lHoP and dlHoP along with the number of false positives per experiment (fp).

	lHoP (199.8 fp)			dlHoP (166.4 fp)		
	cow	horse	back.	cow	horse	back.
cow	32.44	40.36	27.19	62.57	15.0	22.43
horse	12.25	68.63	19.11	6.50	78.45	15.05
AP	0.31			0.51		

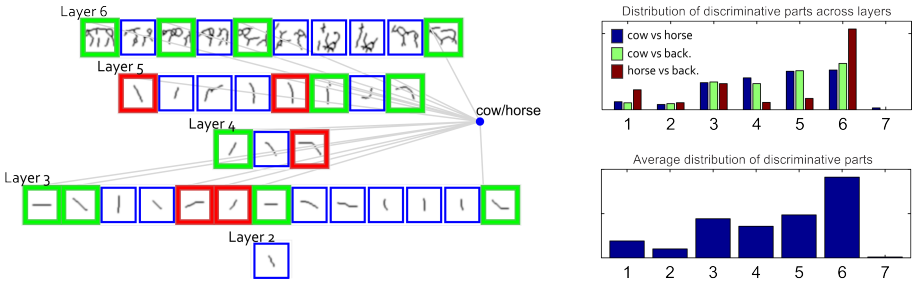

Fig. 4. Distribution of nodes used for discrimination in the Cow-Horse dataset for the cow/horse discriminative node only (selected parts shown in green and red, while the blue parts were selected by the cow/back and horse/back nodes). We also show examples of the frequently selected vocabulary parts by each discriminative node.

4.2 Using discriminative nodes for hypothesis rescoring

We have analyzed the performance of the discriminative nodes on the ETHZ dataset [3]. This dataset contains five categories, in total 255 images with a large intra-class variation. The experiments were performed under an evaluation protocol that is standard for this dataset: A detector for a category was constructed by training from a half of the images of that category and tested on all the remaining images (including all other categories). This experiment was repeated over five random splits of train/test data set. A detection was taken as a true positive if its predicted bounding box intersected with the ground truth by more than 50% and was taken as a false positive otherwise. The lHoP was trained on the training set images and produced a hierarchy with seven layers, together on average resulting in 525 vocabulary parts per experiment. At detection stage, the lHoP and dlHoP processed images at five scales. The positive examples for the dlHoP discriminative nodes were obtained from the lHoP detections on the train images that overlapped with the ground truth by at least 40%. Those detections whose overlap was less than 10% were used for the negative examples (the background category). This was the case in all categories except for the mugs category, since only mugs, whose handle points to the right-hand side, are annotated as ground truth. To avoid using the glasses and mugs with left-pointing handle as negative examples, we have used the Caltech101 background images to obtain the negative examples for the mug category.

Many methods for object detection [19–21, 3] proceed to object detection in several stages. First, hypotheses of the object's location in the image are formed, and then these detections are rescored by assigning each hypothesis an object-certainty value. As a final step they apply hypothesis verification, at which a HOG-like descriptor, combined with a classifier, is used to finely reclassify the area around the predicted, rescored, hypotheses. The first stages depend on the ability of the approach to form good hypotheses, while the last step mainly speaks of performance of the HOG-like descriptors. We apply the dlHoP as a hypothesis forming and rescoring algorithm and we therefore compare its per-

formance to the hypothesis rescoring stages of the lHoP: Each hypothesis from the lHoP (i.e., predicted bounding box) is classified by (6) and then rescored by (5). A common protocol for comparing the effectiveness of the scoring function is to report the recall at 1 false positive per image (FPPI) [19–21, 3]. The results are sumarized in Table 2, where we also report the number of vocabulary nodes selected by the dlHoP at each category. We can see that discriminative nodes significantly improve the lHoP's prediction by rescoring hypotheses and deliver a competitive performance compared to the rescoring stages of the state-of-the-art methods. Note that the methods in the last two columns of Table 2 use a pyramid-match kernel (PMK) [22] with a support-vector machine for hypotheses rescoring. Nevertheless, using only linear combination of discriminative nodes from the lHoP's vocabulary (on average nine per category) delivers competitive results. The lHoP on its own outperforms the related methods that do not apply the PMK hypotheses ranking and produces comparable results to [20] that does use the PMK. The discriminative nodes improve on average the lHoP's performance by five percent and deliver comparable performance to the best hypothesis voting method that applies the PMK [21]. For a further insight we have visualized the distribution of the discriminative vocabulary parts over the layers at one of the runs in Figure 5. We can observe that the discriminative nodes choose the reconstructive vocabulary parts from different layers for each category. Nevertheless, the general trend appears to be selection of most parts from between layers three and five, emphasizing the importance of more global, distinctive, features for reliable categorization.

Table 2. Hypothesis voting and ranking stage detection rates using the Pascal 50% overlap criterion on ETHZ [3] at FPPI=1.0. The N_{disc} denotes the number of discriminative nodes selected by dlHoP along with standard deviation in brackets.

	lHoP [12]	dlHoP [N_{disc}] **our work**	PSM [21]	Hough [3]	w_{ac} [20]	M^2HT [19]	PMK [20]	PMK [21]
Apple	92.5	92.5 [5.2 (1.3)]	90.4	43.0	80.0	85.0	80.0	90.4
Bottle	79.6	85.4 [7.4 (1.7)]	84.4	64.4	92.4	67.0	89.3	96.4
Giraffe	75.1	82.3 [13 (4.6)]	50.0	52.2	36.2	55.0	80.9	78.8
Mug	85.9	86.5 [13.2 (6.9)]	32.3	45.1	47.5	55.0	74.2	61.4
Swan	58.6	70.5 [6 (2.6)]	90.1	62.0	58.8	42.5	68.6	88.6
Average	78.3	**83.4 [9.0 (5.1)]**	69.4	53.3	63.0	60.9	78.6	83.2

5 Conclusion

We have proposed an approach to improve hypotheses rescoring and classification in part-based generative hierarchical compositional models. Our main contribution is the introduction of discriminative nodes – a sparse discriminative network – placed on top of a generative hierarchy. In this network each discriminative node is a linear combination of responses from the activated vocabulary parts

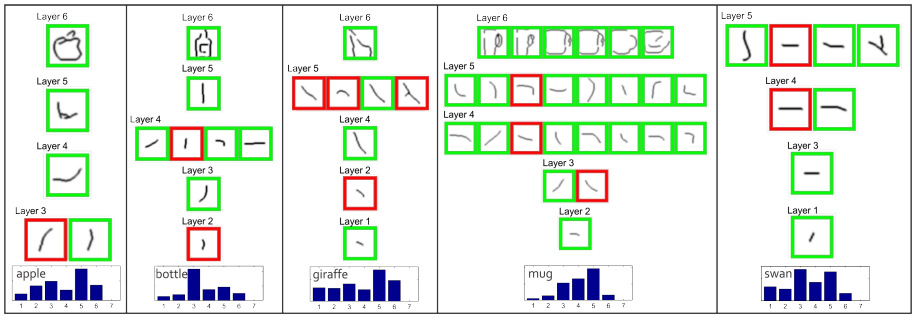

Fig. 5. Distribution of selected vocabulary parts over hierarchy, and examples of the frequently chosen parts in the ETHZ experiment.

at the predicted bounding box. Sparse connectivity is achieved by treating the problem of the discriminative node estimation in the framework of sparse logistic regression. As a proof-of-concept we have applied the methodology to a state-of-the-art hierarchical compositional model lHoP [4]. Experiments show that the discriminative performance of the lHoP consistently and considerably improves with addition of the discriminative nodes. We have observed that the discriminative nodes often chose more complex parts for discrimination, however, the parts are generally dispersed throughout the hierarchy, suggesting utilization of various levels granularity. It is important to note that, our results show that the discriminative parts of the vocabulary occupy only a fraction of an average vocabulary in a generative hierarchy (less than ten percent). These results also speak in favor of using overcomplete, yet sparse, hierarchically organized vocabularies for object detection. On one hand, these vocabularies contain the information used for reconstruction, while at the same time contain the parts relevant for discrimination. An appealing feature of the proposed discriminative nodes is that they do not hamper the scalability and incrementality of the hierarchical model. Note that the application of our methodology is by no means restricted to lHoP, but can easily be applied to any generative hierarchical model.

Acknowledgments. We would like to thank the authors of [4] for providing the code for the lHoP and allowed us using some of their images in our figures. This work was supported in part by ARRS research program P2-0214 and ARRS research projects J2-4284, J2-3607 and J2-2221.

References

1. Opelt, A., Pinz, A., Zisserman, A.: Learning an alphabet of shape and appearance for multi-class object detection. Int. J. Comput. Vision **80**(1) (2008) 16–44
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. **32**(9) (2010) 1627–1645

3. Ferrari, V., Jurie, F., Schmid, C.: From images to shape models for object detection. Int. J. Comput. Vision **87**(3) (2010) 284–303
4. Fidler, S., Leonardis, A.: Towards scalable representations of object categories: Learning a hierarchy of parts. In: Proc. Conf. Comp. Vis. Pattern Recognition. (2007) 1–8
5. Todorovic, S., Ahuja, N.: Learning subcategory relevances for category recognition. In: Proc. Conf. Comp. Vis. Pattern Recognition. (2008) 1–8
6. Zhu, L.L., Chen, Y., Torralba, A., Freeman, W., Yuille, A.: Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. In: Proc. Conf. Comp. Vis. Pattern Recognition. (2010)
7. Kokkinos, I., Yuille, A.: Inference and learning with hierarchical shape models. Int. J. Comput. Vision **93**(3) (2011) 1–25
8. Si, Z., Zhu, S.: Unsupervised learning of stochastic and-or templates. In: Int'l Workshop on Stochastic Image Grammar. (2011)
9. Poon, H., Domingos, P.: Sum-product networks: A new deep architecture. In: Proc. 12th Conf. on Uncertainty in Artificial Intelligence. (2011) 337–346
10. Lee, H., Grosse, R., Ranganath, R., Ng, A.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proc. Int. Conf. Mach. Learning. (2009) 609–616
11. Kavukcuoglu, K., Sermanet, P., Boureau, Y., Gregor, K., Mathieu, M., LeCun, Y.: Learning convolutional feature hierarchies for visual recognition. In: Neural Inf. Proc. Systems. (2010)
12. Fidler, S., Boben, M., Leonardis, A.: Evaluating multi-class learning strategies in a hierarchical framework for object detection. In: Neural Inf. Proc. Systems. (2009)
13. Epshtein, B., Ullman, S.: Satellite features for the classification of visually similar classes. In: Proc. Conf. Comp. Vis. Pattern Recognition. Volume 2. (2006) 2079–2086
14. Yamashita, O., Sato, M., Yoshioka, T., Tong, F., Kamitani, Y.: Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage **42**(4) (2008) 1414–1429
15. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Discriminative learned dictionaries for local image analysis. In: Proc. Conf. Comp. Vis. Pattern Recognition. (2008) 1–8
16. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th Int'l Conf. Computer Vision. Volume 2. (July 2001) 416–423
17. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. Int. J. Comput. Vision **77**(1) (2008) 259–289
18. Borenstein, E., Ullman, S.: Combined top-down/bottom-up segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2) (2008) 2109 – 2125
19. Maji, S., Malik, J.: Object detection using a max-margin hough transform. In: Proc. Conf. Comp. Vis. Pattern Recognition. (2009) 1038–1045
20. Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. In: Proc. Int. Conf. Computer Vision. (2009) 484–491
21. Riemenschneider, H., Donoser, M., Bischof, H.: Using partial edge contour matches for efficient object category localization. In: Proc. European Conf. Computer Vision. (2010) 29–42
22. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. The Journal of Machine Learning Research **8** (2007) 725–760