Automatic Discovery of Meaningful Object Parts with Latent CRFs

Paul Schnitzspan[1] Stefan Roth[1] Bernt Schiele[1,2]
[1]Department of Computer Science, TU Darmstadt [2]MPI Informatics, Saarbrücken

Abstract

Object recognition is challenging due to high intra-class variability caused, e.g., by articulation, viewpoint changes, and partial occlusion. Successful methods need to strike a balance between being flexible enough to model such variation and discriminative enough to detect objects in cluttered, real world scenes. Motivated by these challenges we propose a latent conditional random field (CRF) based on a flexible assembly of parts. By modeling part labels as hidden nodes and developing an EM algorithm for learning from class labels alone, this new approach enables the automatic discovery of semantically meaningful object part representations. To increase the flexibility and expressiveness of the model, we learn the pairwise structure of the underlying graphical model at the level of object part interactions. Efficient gradient-based techniques are used to estimate the structure of the domain of interest and carried forward to the multi-label or object part case. Our experiments illustrate the meaningfulness of the discovered parts and demonstrate state-of-the-art performance of the approach.

1. Introduction

It has long been argued that part-based models [12] are a powerful paradigm for object class recognition, due to both their expressiveness and intuitive interpretation. Consequently, many part-based approaches have been proposed [1, 2, 3, 7, 9, 11, 19, 25] showing favorable properties such as robustness to partial occlusion [19] and articulation [1, 9], and the ability to deal with viewpoint changes [2]. Part-based models are often factorized into different components [9, 11], which can enable knowledge transfer across classes [28, 29]. Part representations can moreover be integrated hierarchically [2, 7] to improve interpretability and performance.

At the same time, global template models [4, 13, 18, 31] have been highly competitive on difficult datasets, typically by combining high-dimensional feature vectors with powerful discriminative classifiers. Despite their impressive performance, these methods often suffer in the presence of partial occlusion and articulation. To combine the advantages of both worlds, global templates have been enriched with the notion of parts [5, 10]. In this paper we go beyond previous work by learning meaningful parts, their spatial extent, and their topological structure in a fully automatic fashion.

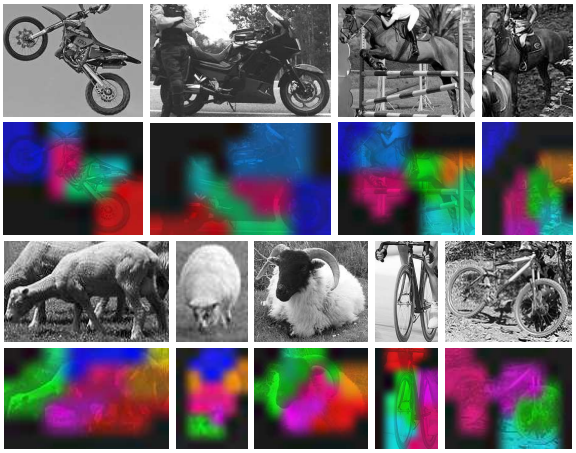

Figure 1. Parts of motorbikes, horses, bikes and sheep automatically discovered by our approach. Note how different viewpoints, articulation, and partial occlusions can be handled.

The goal of this paper is to introduce a novel model for object classes that brings together the competitive power of discriminative learning with the flexibility and expressiveness of part-based models. An important consideration is that we do not want to provide supervision at the part level, but instead train the model in a weakly supervised fashion from class labels alone (c.f. [3, 10]). In order to enable the automatic discovery of semantically meaningful part representations of objects, we model part labels as hidden nodes in a graphical model. We rely on two major components: A multi-label conditional random field (CRF) that aggregates image evidence and predicts object part occurrences, and a probabilistic classifier that predicts object or background occurrence from the spatial part configuration (see Fig. 2). In order to avoid having to provide part labels, we take a Bayesian approach and marginalize out the latent part configuration. Thus, our detector is a mixture of part-driven classifiers, which can take advantage of the uncertainty of bottom-up part discovery. Fig. 1 shows examples of how our approach automatically discovers semantically meaningful parts for PASCAL VOC 2007 motorbikes, horses, bikes and sheep. Note how the part interpretation stays consistent across different viewpoints and other intra-class variations, and how articulation and partial occlusions are detected and handled. To train both the part

CRF as well as the part-driven object classifier, we develop an expectation maximization (EM) algorithm that only requires bounding box labels as given in many object detection datasets. Our experimental results demonstrate that our model not only enables learning of semantically meaningful parts, but obtains competitive results on the difficult PASCAL VOC 2007 dataset.

Related work. Generative part-based approaches [3, 9, 11, 19, 29] have been shown to achieve high flexibility by adapting a factorizable composition of object parts. These components can be interpreted in a semantically meaningful way, as they often reoccur across object instances. Utilizing object parts has shown to be beneficial, especially in the presence of partial occlusion and articulation. Nonetheless, such approaches often cannot compete with discriminatively trained part representations [5, 6, 10, 16, 17, 20, 27, 30] on benchmark datasets.

Discriminative part-based models [5, 6, 10, 20, 27] enrich a global object template with a notion of parts that increases the performance, especially for difficult datasets with few training samples. In contrast to [5, 10], we consider parts with flexible shapes and sizes, and marginalize over part occurrences instead of maximization. This allows us to take advantage of the inherent uncertainty of bottom-up part prediction, since we do not assume all parts to be present (see Fig. 1). Note here that marginalizing (i.e., summing probabilities) differs from summing scores (i.e., log-probabilities). Moreover, [10] relies on a star-shaped part constellation, which allows parts to occur in a local neighborhood of a canonical rigid representation. We go beyond this relatively restricted model of part configurations, and allow for and learn arbitrary part topologies. [6, 20] propose to boost multiple instances of part occurrences in order to provide a flexible model that robustly copes with partial occlusion and articulation. We extend these approaches by automatically learning the structure of the domain of interest, which in turn allows to directly model the dependencies of object parts within the spatial extent of the object. [27] estimates a part labeling using k-means clustering of the features and uses it as pseudo-data for learning. Our model is much more flexible by treating the part labels as hidden nodes in order to adapt to the best constellation.

Hidden conditional random fields [16, 23] provide expressive models for gesture and object recognition. By treating part labels as hidden variables they combine the expressive power of latent models with discriminative modeling, and provide the flexibility of adapting to the best constellation of parts. But while hidden CRFs are powerful, the object class prediction depends on the part labeling in a (sometimes fully) factorized fashion. Our latent CRF extends these approaches by separating the bottom-up part prediction and the object classifier, and uses a robust summation-based classifier that is tolerant to missing parts, as they can occur under partial occlusion. To increase the flexibility and expressiveness of CRFs, structure learning in graphical models has been recently introduced [24, 26]. We extend the work of [26] from the binary object vs. background case to the multi-label case and learn pairwise couplings of a latent part representation in a weakly supervised setting using EM. Multi-label structure learning increases the semantic interpretability of inferred part constellations.

Multi-feature approaches [14, 31] have shown impressive levels of performance on difficult object detection datasets, which inspired us to keep our model general by allowing for combinations of orthogonal feature descriptors. While we use only two different feature types in this paper, such multi-feature approaches are complementary and should enable further performance improvements.

2. Latent CRF model

In our object detection scenario we are given a set of M images (or more precisely bounding boxes) $X = (\mathbf{x}^1, \dots, \mathbf{x}^M)$, which contains objects from a particular class and background images. We are also given observed variables $Y = (y^1, \dots, y^M)$, $y^m \in \{0, 1\}$, which specify whether the corresponding image contains the object of interest ($y^m = 1$) or not ($y^m = 0$). We additionally introduce latent variables $Z = (\mathbf{z}^1, \dots, \mathbf{z}^M)$, which refer to inferred part labelings for each image. Every part labeling $\mathbf{z}^m$ consists of N variables $\mathbf{z}^m = (z_1^m, \dots, z_N^m)$, which denote localized part labels and are represented by the output nodes of the part CRF. Each $z_i^m \in \{0, \dots, P\}$ takes on one of the possible part labels. P refers to the (maximum) number of object parts; part 0 represents the background. Without yet specifying the CRF in detail, we denote its nodes as $V = (1, \dots, N)$ and let E refer to the edges of the graph. For simplicity of notation, we drop the superscript m indicating the training instance wherever applicable.

Since in object detection we are interested in the probability of presence or absence of objects, we model the posterior directly by marginalizing out the latent variables $\mathbf{z}$:

$$p(y|\,\mathbf{x}; \theta, E) = \sum_{\mathbf{z}} \underbrace{p(y|\mathbf{z}; \boldsymbol{\gamma})}_{\text{part-driven classifier}} \underbrace{p(\mathbf{z}|\mathbf{x}; \boldsymbol{\alpha}, \mathbf{e}, E)}_{\text{part CRF}}, \quad (1)$$

where the set of parameters is given by $\theta = \{\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{e}\}$. Here we assume that $p(y|\mathbf{z}; \boldsymbol{\gamma})$ is conditionally independent of $\mathbf{x}$ given $\mathbf{z}$, which implies that the object classifier only relies on the inferred part configuration rather than on the image itself. The part CRF models the distribution of object parts, and by marginalizing over $\mathbf{z}$ we obtain a mixture of part-driven object classifiers (see Fig. 2). The marginalization has the advantage that rather than committing to a possibly wrong configuration early on and drawing the wrong conclusion in consequence, we can consider all possible part configurations and take advantage of the inherent uncertainty of bottom-up part prediction.

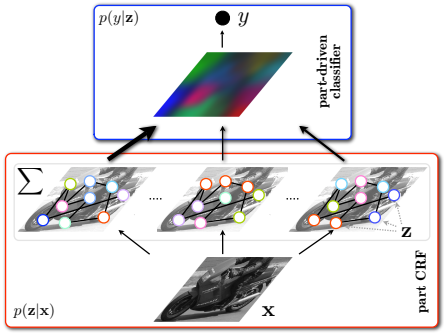

Figure 2. Model architecture consisting of a part CRF for bottom-up part detection, and part-driven object classifier. The part variables are marginalized out, taking advantage of their uncertainty.

2.1. Part CRF

We build our part model on CRFs since they provide a direct way of using multiple labels to represent spatially distributed parts in an image, and allow to model the uncertainty of the part configuration. The nodes in our graphical model are distributed over the image plane in an arbitrary layout, and are linked to features computed from certain image regions. We associate a part label with each node, no matter where the corresponding feature is located or how small or large the feature size relative to the bounding box is. Several nodes can be associated with the same part label, thus allowing our model to flexibly adapt the spatial extent of object parts. During learning, the spatial extent of the object parts, their hierarchical feature representation, as well as the graph structure (= object topology) are determined automatically. In that sense our model is more powerful and far more flexible than previous work [10, 16, 17, 27].

Through the connection of nodes to image regions, the assigned node labels can be interpreted as a semantic representation of localized object parts. To deal with this flexible domain, we automatically learn the pairwise structure of feature couplings to allow arbitrary part interactions within the spatial extent of objects. The posterior distribution $p(\mathbf{z}|\mathbf{x})$ of part labels $\mathbf{z}$ given an image $\mathbf{x}$ is modeled as a CRF with unary and pairwise potentials:

$$p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E) = \frac{1}{\mathcal{Z}(\boldsymbol{\alpha},\mathbf{e},\mathbf{x},E)} \prod_{i\in V} \psi_i(z_i,\mathbf{x};\boldsymbol{\alpha}) \cdot \prod_{(i,j)\in E} \phi_{ij}(z_i,z_j,\mathbf{x};\mathbf{e}). \quad (2)$$

The unaries ψ_i aggregate part evidence from a single image feature, while the pairwise potentials ψ_{ij} allow taking advantage of pairwise feature couplings. $\mathcal{Z}(\boldsymbol{\alpha},\mathbf{e},\mathbf{x},E)$ is the partition function, which ensures normalization. For now we assume that the graph structure is given, i.e. the set of edges $E \subseteq V \times V$ is fixed, and describe in Sec. 3 how the graph structure is learned as well.

Unary potentials. The unary potentials in our model are responsible for modeling part occurrences in an image based on single features $f_i(\mathbf{x})$. To supply robust discriminative part classifiers we train one support vector machine (SVM) $F(\cdot,\cdot)$ per object part and feature type. For now we assume that a part assignment is given for SVM training; in Sec. 3 we describe in detail how the part assignment is estimated from the class labels. Note that the part classifier is shared between all possible locations of a localized feature, which allows the object parts to occur anywhere in the bounding box and enables the model to capture articulations, viewpoint changes, and partial occlusions. In contrast to global object descriptors, this allows the representation to more easily adapt to positional variations and enables more specific appearance models. Based on the SVM classifier, we define the unary potential of node i for part label z_i using a softmax as

$$\psi_i(z_i,\mathbf{x};\boldsymbol{\alpha}) = \frac{\exp\left(F(\boldsymbol{\alpha}(z_i), f_i(\mathbf{x}))\right)}{\sum_{c=0}^{P} \exp\left(F(\boldsymbol{\alpha}(c), f_i(\mathbf{x}))\right)}, \quad (3)$$

where the $f_i(\mathbf{x})$ refer to the features as described in Sec. 4, and $\boldsymbol{\alpha}(z_i)$ denotes the support vector coefficients of part label (class) z_i. The background "part" is modeled as $F(\boldsymbol{\alpha}(0), f_i(\mathbf{x})) = const$.

Pairwise potentials. The pairwise potentials capture the structure of the domain of interest by modeling the co-occurrence of parts at connected nodes (i,j). Based on the interaction of the corresponding image features $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$, we capture the interplay of parts by computing softmax classifiers on the concatenated features:

$$\phi(z_i,z_j,\mathbf{x};\mathbf{e}) = \frac{\exp\left((f_i(\mathbf{x}),f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{z_i z_j}\right)}{\sum_{c_1,c_2=0}^{P} \exp\left((f_i(\mathbf{x}),f_j(\mathbf{x}))^T \mathbf{e}_{ij}^{c_1 c_2}\right)} \quad (4)$$

Here, the parameters $\mathbf{e}_{ij}^{z_i z_j}$ are specific to each pairwise coupling and each combination of part labels. By allowing connections between arbitrary pairs of nodes, we obtain the flexibility to represent spatial relations not only of object parts in local neighborhoods, but also between distant locations within the spatial extent of objects. This topology is more flexible than simple star-shaped part models [10].

2.2. Part-driven object classifier

Given a spatial distribution of object parts the part-driven object classifier estimates object or background occurrence. We set up this object classifier in a non-parametric way that allows to model the contribution of each localized part label toward the object or background hypothesis. This holistic interpretation of part occurrences has the advantage that it allows for ambiguities in part localization as well as in part annotation. Such ambiguities are inevitable particularly in our latent variable setting, in which parts are inferred auto-

matically. The object classifier can be written as

$$p(y|\mathbf{z};\boldsymbol{\gamma}) = \begin{cases} \sum_{i\in V} \gamma_i(z_i), & y = 1 \\ 1 - \sum_{i\in V} \gamma_i(z_i), & y = 0 \end{cases} \quad (5)$$

with $\sum_{i\in V}\sum_{c=0}^{P} \gamma_i(c) = 1$ to ensure normalization. Note that by defining the classifier through weighted sums of part occurrences, it remains robust to an occasional absence of parts. This is in contrast to hidden CRFs that instead assume a factorized model [16, 23]. During training, as will be explained in Sec. 3, the parameters $\boldsymbol{\gamma}$ are learned from the inferred part occurrences in the training set.

2.3. Detecting object instances

In order to evaluate whether an object instance is present in the bounding box $\mathbf{x}$ or not, we need to compute $p(y = 1|\mathbf{x};\theta, E) = \sum_{\mathbf{z}} p(y = 1|\,\mathbf{z};\boldsymbol{\gamma})p(\mathbf{z}\,|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e}, E)$, which involves marginalization over all part configurations. As derived in the supplementary material, the posterior object probability simplifies to

$$p(y = 1|\mathbf{x};\theta, E) = \sum_{i\in V}\sum_{z_i=0}^{P} \gamma_i(z_i)p(z_i|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e}, E). \quad (6)$$

Since exact computation of the marginals $p(z_i|\,\mathbf{x};\boldsymbol{\alpha},\mathbf{e}, E)$ is intractable, we approximate them using the beliefs $b_i(z_i)$ from sum-product belief propagation. For efficiency reasons, we pre-filter candidate bounding boxes with a HOG detector [4] and compute the score of the pre-filtered windows with our full model. For the sake of completeness we also report the performance without pre-filtering.

3. Learning the Model

To train our latent-variable model, we rely on the well-known expectation maximization (EM) algorithm. This allows our approach to discover semantically interpretable part annotations from object class labels alone. The combination of the flexibility of our model and the power of EM to infer and adapt to soft assignments of hidden nodes, and therefore part labels, yields a theoretically sound, yet practically scalable approach. Our objective is to maximize the expected complete log-likelihood w.r.t. $\theta = \{\boldsymbol{\gamma},\boldsymbol{\alpha},\mathbf{e}\}$:

$$Q(\theta,\theta^{\text{old}}) = \sum_{Z} \Big[p(Z|Y, X;\theta^{\text{old}}, E^{\text{old}}) \cdot \log\big(p(Y|Z;\boldsymbol{\gamma})p(Z|X;\boldsymbol{\alpha},\mathbf{e}, E)\big)\Big], \quad (7)$$

where θ^{old} refers to the parameters from the last M-step.

Initialization. It is well known that EM requires proper initialization to work well. We infer an initial part labeling using k-means clustering over the positive training instances, which yields a hard assignment of nodes to object parts (c.f. [27]). We use these hard part assignments from the positive instances and randomly sampled negative instances to initially train the part classifiers, i.e. SVMs, which yields our initialization for $\boldsymbol{\alpha}$. Note here that this procedure only provides an initialization; later the part classifiers are retrained as required by the part representation. The parameters of the part-driven classifier $\boldsymbol{\gamma}$ are initialized by counting part occurrences in the inferred hard assignment and normalizing. The edge parameters $\mathbf{e}$ require no initialization, as at the beginning no edges are present in the graph (see below).

E-Step. In the E-step we compute expected (i.e. soft) assignments of part labels to nodes for the training set of observed class labels Y and images X:

$$p(Z|Y, X;\theta^{\text{old}}, E^{\text{old}}) = \prod_{m=1}^{M} \frac{p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}^{\text{old}})p(\mathbf{z}^m|\mathbf{x}^m;\boldsymbol{\alpha}^{\text{old}},\mathbf{e}^{\text{old}}, E^{\text{old}})}{p(y^m|\mathbf{x}^m;\theta^{\text{old}}, E^{\text{old}})} \quad (8)$$

Here, $p(y^m|\mathbf{x}^m;\theta^{\text{old}}, E^{\text{old}})$ is computed approximately using belief propagation as in Eq. (6). The E-step thus yields probabilities of nodes belonging to certain object parts, which flexibly adapts toward a meaningful representation of parts as learning proceeds.

Generalized M-Step. After computing the soft part assignments, we maximize the expected complete log-likelihood $Q(\theta,\theta^{\text{old}})$ w.r.t. θ. We use gradient ascent, since there is no closed form solution for the parameters $\boldsymbol{\gamma}$. The gradient can be approximated as[1]

$$\frac{\partial Q(\theta,\theta^{\text{old}})}{\partial\gamma_i(c)} \approx \sum\nolimits_{m=1,y^m=1}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m;\theta^{\text{old}}, E^{\text{old}})} - \sum\nolimits_{m=1,y^m=0}^{M} \frac{b_i^{\text{old}}(c)}{p(y^m|\mathbf{x}^m;\theta^{\text{old}}, E^{\text{old}})}, \quad (9)$$

where b_i^{old} are the part beliefs from the previous iteration. We use one gradient update per M-step.

We optimize the parameters $\boldsymbol{\alpha}$ by retraining the SVM part classifier. To exploit the uncertainty of the part labeling expressed by the soft assignments, we weigh each part occurrence in the training set with its soft assignment. Here we alternate between the max-margin objective of SVM training and likelihood maximization. In future work we will investigate how to overcome this potential restriction and study how to learn all parameters in a single formalism.

As is the case even for fully observed training of CRFs, there is no closed form solution for the edge parameters $\mathbf{e}$. We thus use gradient ascent that locally maximizes $Q(\theta,\theta^{\text{old}})$. The gradient with respect to edge parameters

[1] This and the following equations are derived in the suppl. material.

for parts c_1 and c_2 is given as

$$\frac{\partial Q(\theta,\theta^{\text{old}})}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = \sum_{m=1}^{M}\left[\sum_{\mathbf{z}^m}\left(\frac{\partial \log(p(\mathbf{z}^m|\mathbf{x}^m,\boldsymbol{\alpha},\mathbf{e},E))}{\partial \mathbf{e}_{ij}^{c_1 c_2}} p(y^m|\mathbf{z}^m;\boldsymbol{\gamma}^{\text{old}})\right.\right.$$
$$\left.\left. p(\mathbf{z}^m|\mathbf{x}^m,\boldsymbol{\alpha}^{\text{old}},\mathbf{e}^{\text{old}},E^{\text{old}})/p(y^m|\mathbf{x}^m,\theta^{\text{old}},E^{\text{old}})\right)\right]. \quad (10)$$

Computing the gradient of $Q(\theta,\theta^{\text{old}})$ thus reduces to calculating the conditional log-likelihood gradient, which is also required for training standard CRFs. The gradient of the conditional log-likelihood $\mathcal{C}(\mathbf{e}) = \log p(\mathbf{z}\,|\,\mathbf{x},\boldsymbol{\alpha},\mathbf{e},E)$ with respect to the edge parameters is given as:

$$\frac{\partial \mathcal{C}(\mathbf{e})}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = E_{\mathbf{z}_{\{z_i=c_1,z_j=c_2\}}|\mathbf{x}}\left[(f_i(\mathbf{x}),f_j(\mathbf{x}))^T\phi_{ij}(c_1,c_2,\mathbf{x})\right] - E_{p(\mathbf{z}_{\{z_i=c_1,z_j=c_2\}}|\mathbf{x})}\left[(f_i(\mathbf{x}),f_j(\mathbf{x}))^T\phi_{ij}(c_1,c_2,\mathbf{x})\right], \quad (11)$$

where $E_{\mathbf{z}|\mathbf{x}}$ denotes the empirical expectation and $E_{p(\mathbf{z}|\mathbf{x})}$ refers to the expectation under the posterior distribution of our part CRF (see e.g. [26]).

In order to cope with differing major orientations of the instances (left-right), we initialize our model with the original orientations of the dataset. In each iteration we evaluate our model on the original and a mirrored image and choose the one with the highest score for the next iteration.

3.1. Structure learning

In order to learn the spatial relationship between object parts, we use discriminative structure learning to find the edges in the CRF that maximize the discriminative power of the overall model. In particular, we use L1-regularized gradient-based discriminative structure learning [24, 26] and extend it to the multi-label case. In our scenario we are interested in improving the discriminative power of our approach in terms of the object's presence or absence, and therefore consider the log-posterior ratio

$$\mathcal{R}(\mathbf{e}_{ij}) = \max_{(c_1,c_2)\in\{0..P\}^2}\left\|\sum_{m=1,y^m=1}^{M}\frac{\partial \log p(y^m=1|\mathbf{x}^m;\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} - \sum_{m=1,y^m=0}^{M}\frac{\partial \log p(y^m=0|\mathbf{x}^m;\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}}\right\|. \quad (12)$$

The edges that maximize this ratio likely improve our model, and therefore should be added to it. At the same time, edges that have small ratio and small absolute edge weights can be removed from the current active edge set, because they have only a small impact on the objective. In contrast to [26] the structure learning objective $\mathcal{R}(\mathbf{e}_{ij})$ considers the gradient over all possible edges and possible part constellations and takes the maximum. After some algebraic reformulations we can write:

$$\frac{\partial \log p(y|\,\mathbf{x};\theta,E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}} = \frac{1}{p(y|\mathbf{x};\theta,E)} \left(\sum_{\mathbf{z}} p(y|\mathbf{z};\boldsymbol{\gamma})p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)\frac{\partial \log p(\mathbf{z}|\mathbf{x};\boldsymbol{\alpha},\mathbf{e},E)}{\partial \mathbf{e}_{ij}^{c_1 c_2}}\right). \quad (13)$$

The gradient on the right hand side is computed as in Eq. (11). We proceed by finding the best edge to add:

$$\mathbf{e}_{i^*j^*} = \underset{(i,j)\in V\times V\setminus E}{\arg\max}\ \mathcal{R}(\mathbf{e}_{ij}) \quad (14)$$

We start the learning process with no pairwise couplings of nodes and iteratively add the ten best edges (highest ratio of gradient norm) to the model at the end of each M-step. At the same time, we remove edges with absolute weight below a threshold τ_1 that also have an absolute gradient norm below the threshold τ_2. In combination with L1-regularization, this scheme leads to sparsely connected graphs, and at convergence has a connectedness of ~20%. Experiments with different τ_1 and τ_2 showed that the determined structure is robust to changes in τ_1 and τ_2 even though these parameters control the connectedness of our model – higher thresholds yield lower connectedness.

4. Image Features

The aggregation of features and their linkage to image regions defines the basic spatial layout of the latent nodes of our model. We build a dense representation of objects that includes both histograms of oriented gradients (HOG) [4] and bag-of-words (BoW) [18] descriptors. In our experiments these specific feature descriptors emerged to be suitable to capture local deformations and viewpoint changes. Note, however, that our model allows for an arbitrary layout of nodes, which means that we can rely on any feature aggregation scheme, both local and global. This allows for future integration of orthogonal features, which appears promising due to the success of combining several features [31]. In our experiments we fixed the extent of our features (the size of the linked image region) and leave it for future work to automatically select optimal feature scopes.

HOG descriptors. The HOG descriptors are computed by calculating a dense grid of non-overlapping cells of oriented gradients [4] - each cell being 8×8 pixels in size. A dense block grid with 50% overlap between blocks is built by concatenating and normalizing four neighboring cells. Similar to [10], we rely on local views of objects by concatenating several neighboring blocks to form one feature descriptor. In our experiments we concatenate 5×5 neighboring blocks into one local descriptor. In addition, we compute a global descriptor that comprises all blocks of the grid, thus aggregating evidence from the entire object.

BoW descriptors. The bag of words (BoW) descriptors [18] are formed by densely calculating SIFT features [21]

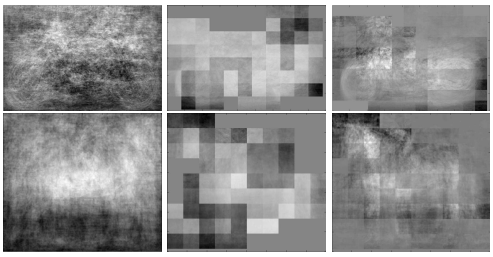

Figure 3. Mean image averaged over *(left)* all instances, *(middle)* part occurrences of k-means clustering, *(right)* part occurrences learned with EM. *(top)* VOC 2007 motorbikes, *(bottom)* horses.

with radii $(5, 10, 15)$ and a spacing of 10 pixels. We vector-quantize these features with k-means clustering over the positive training instances. We divide the image into overlapping regions, which each forms a feature descriptor. In our experiments we use regions of 50×50 pixels with an overlap of 50%. A local BoW descriptor is then formed by measuring the word occurrences in one specific region. A global BoW descriptor is calculated by measuring the word occurrences in the entire bounding box.

Part classifiers. For each local as well as global feature and feature type $\mathcal{T} \in \{\mathcal{H}, \mathcal{B}\}$ ($\mathcal{H}$ denotes HOG, and $\mathcal{B}$ BOW) we train one SVM

$$F^{\mathcal{T}}(\boldsymbol{\alpha}^{\mathcal{T}}(c), f_i^{\mathcal{T}}(\mathbf{x})) = \sum_{\mathbf{s} \in S^{\mathcal{T}}(c)} \alpha_s^{\mathcal{T}}(c) K\left(\mathbf{s}, f_i^{\mathcal{T}}(\mathbf{x})\right) + \alpha_0^{\mathcal{T}}(c), \tag{15}$$

where $S^{\mathcal{T}}(\cdot)$ refers to the set of support vectors and $K(\cdot, \cdot)$ denotes an appropriate Mercer kernel. In our experiments we make use of the histogram intersection kernel [22]. Each part classifier is then defined as a sum of HOG and BoW classifiers $F(\boldsymbol{\alpha}(c), f_i(\mathbf{x})) = F^{\mathcal{H}}(\boldsymbol{\alpha}^{\mathcal{H}}(c), f_i^{\mathcal{H}}(\mathbf{x})) + F^{\mathcal{B}}(\boldsymbol{\alpha}^{\mathcal{B}}(c), f_i^{\mathcal{B}}(\mathbf{x}))$.

5. Experiments

We evaluated our model on the PASCAL VOC 2007 dataset [8] with the common average precision (AP) metric. This dataset includes images from 20 object classes and is challenging due to partial occlusion, articulation, and viewpoint changes. Training our model takes ~8 hours while computing detection scores for an entire image takes ~15 sec on a 2 GHz AMD Opteron machine, when using a global HOG pre-filter. Without the pre-filter we apply belief propagation for all locations and scales, which increases the computation time to ~450 sec per image. We use *SVM*$^{\text{light}}$ [15] for training the SVMs.

Qualitative observations. Fig. 3 shows mean images over the positive training instances of motorbikes (top) and horses (bottom). The left column shows the global mean image over all bounding boxes, where it is challenging to see any real object class structure. The middle column shows mean part occurrences of a fixed k-means part assignment (c.f. [27]) weighted by their probability and shifted to their canonical location. Even though one can recognize a trend towards the discovery of object parts, the object structure is still rather weak. The right column shows part occurrences of our latent CRF model, where the parts are weighted by their probability from the part CRF and shifted to their canonical location (the spatial mean of the classifier weights $\gamma_i(c)$). It becomes clearly visible that our model automatically discovers object parts, such as wheels of motorbikes or the head of horses, allowing for a much better alignment of instances and parts.

VOC 2007	fixed structure ("no sl")	structure learning ("sl")
global	30.2	-
k-means	32.1	33.2
maximization	33.4	35.0
marginalization	34.2	36.3

Table 1. Comparison of different model instantiations on a subset of PASCAL VOC 2007 motorbikes (in average precision).

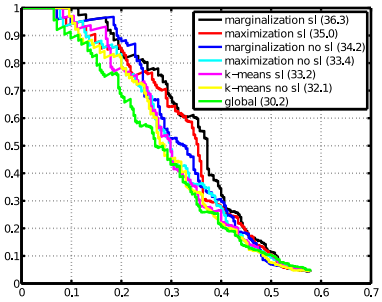

Figure 4. Evaluation of different instantiations of our model on a subset of PASCAL VOC 2007 motorbikes.

Figs. 1, 5, and 6 show object segmentations, where the color-coded part labelings are automatically inferred by our model. The color saturation encodes the probability of each part. As can be seen in Fig. 5 (motorbikes) our model is capable of handling viewpoint variation (row one) as well as partial occlusion (row two left and row three left). These segmentations illustrate one major benefit of our framework: Our model implicitly handles partial occlusions by considering all possible configurations simultaneously and weighing them according to their probability. This avoids relying on the most probable and possibly misleading part labeling. For articulated object classes like horses (Fig. 6) we can observe the same. Our model captures articulation (row one left and row three left), viewpoint variation (row four left) as well as partial occlusion (row two right, row four left). Note how our model adapts to a meaningful representation of parts even for articulated object classes like horses.

Quantitative evaluation on VOC 2007 motorbikes. In Tab. 1 and Fig. 4 we compare different components and settings of our model on a subset (left and right facing) of the images of the motorbike class of the PASCAL VOC 2007 challenge. We show the performance of (i) using only global part descriptors, (ii) using a fixed k-means assign-

ment of parts, (iii) using the most probable part (MAP) per node instead of marginalization, and (iv) using marginalization. All part-based settings are evaluated with a fixed and a learned graph structure. The fixed structure accounts for local neighborhood dependencies that connect each node to its four neighbors in a regular grid, as in standard CRFs.

As can be seen, our full model outperforms the global template model by 6.1% AP, which emphasizes the importance of enriching global models with a semantically meaningful notion of parts. Moreover, treating part labels as hidden nodes is clearly advantageous to fixing them based on k-means clustering (AP increase of 3.1%). This holds true for the case of fixed graph structure as well (gain of 2.1% AP), which shows that the higher expressiveness of latent models results in superior performance. This quantitative evaluation is consistent with our qualitative observations, where parts inferred by our model showed a much better alignment of instances than the k-means instantiation.

In order to show the benefit of marginalizing out all possible part configurations, we compare the marginalization scheme against considering only the maximum part assignment (MAP) for each node. Marginalization shows a gain of 1.3% AP over maximization, which emphasizes the benefit of considering all possible part configurations instead of relying only on possibly misleading maximal responses. Note that structure learning always led to better results than a fixed graph structure, which demonstrates the increased flexibility of the learned structure.

Quantitative evaluation on all VOC 2007 classes. In order to further evaluate the contribution of our work, we show results of different instantiations of our model (using only global parts, using our full model but only HOG features with and without pre-filter, and our full model with HOG and BoW descriptors). Tab. 5 compares those with state-of-the-art approaches. As can be seen our model achieves competitive performance (28.7% AP on average).

Using only HOG features allows a fair comparison to [10], who use similar features. On average over all classes, our flexible part-based approach shows an improvement over [10] of 1.0% AP. We achieve better results on 16 of 20 classes, which emphasizes the benefits of the flexible object topology and marginalizing over all part constellations. Note, that our model without applying the pre-filter (27.1% AP on average across classes) is on par or slightly better than inferring our model on pre-filtered hypotheses (26.9% AP on average across classes).

We achieve an improvement of 1.5% AP and better results on 17 of 20 classes compared to [5], who considered detections of all object classes within an image by simultaneously inferring a notion of multi-class layout and context. Such additional information is orthogonal to our model and is likely to improve the performance further.

Compared to the structure learning approach of [26], which used similar features but no parts, yields an improvement of 1.2% AP, which shows the advantage of integrating part labels in a structure learning framework.

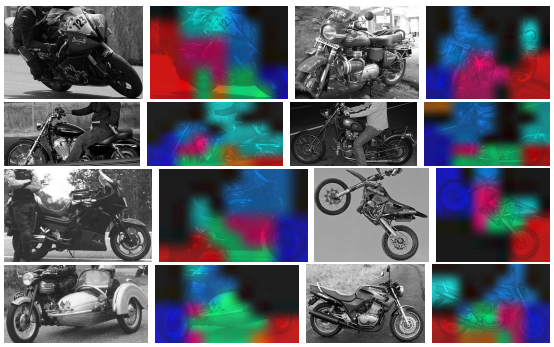

Figure 5. Motorbike segmentation examples (see text for details).

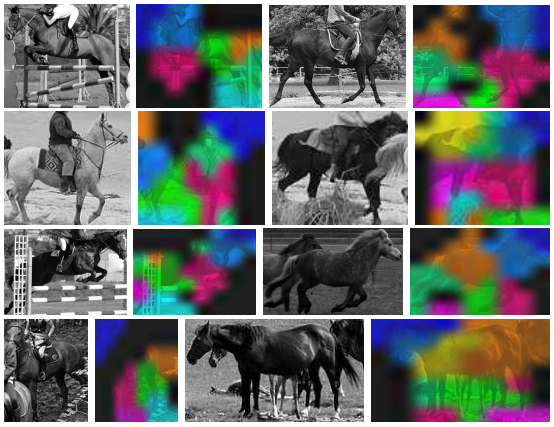

Figure 6. Horse segmentation examples (see text for details).

We achieve better performance than [31] on 5 object categories, even though the latter approach gives better performance on average. It is likely that a large part of this increased performance is due to integrating more complementary feature descriptors, which could also be done in our model as sketched in Sec. 4. Since our model remains general and allows for integration of more features, we expect a substantial performance gain by doing so.

Compared to the original VOC 2007 challenge we achieve a performance gain of 5.4% AP and show better results on 17 of 20 object classes. Here we compare against the best method per class and not against a single model.

6. Conclusions

We presented a novel discriminative framework that successfully combines powerful discriminative learning techniques with the flexibility and expressiveness of part-based models and discriminative pairwise structure learning. We relied on weakly supervised training by treating part labels as hidden nodes, and letting our approach automatically discover semantically meaningful part representations. Our model lends itself to modeling the spatial layout of objects even in the presence of heavy articulation and view-

VOC 2007	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Our model	31.9	57.0	9.1	15.2	26.0	42.7	49.3	14.5	15.2	18.5
Our model (HOG only)	29.1	56.4	4.6	13.0	25.2	40.7	47.3	13.5	10.1	18.8
Our model (HOG only) no pre-filter	28.2	54.8	8.9	10.8	27.2	42.5	45.9	12.2	8.0	20.1
Our model (global)	20.1	45.5	1.8	9.0	19.3	34.6	36.6	11.5	7.3	13.0
DPM (VOC07) [10]	20.6	36.9	9.3	9.4	21.4	23.2	34.6	9.8	12.8	14.0
DPM [10]	28.1	55.4	1.4	14.5	25.4	38.9	46.6	14.3	9.4	16.0
Multi-class layout [5]	28.8	56.2	3.2	14.2	29.4	38.7	48.7	12.4	16.0	17.7
Bin. struct learning [26]	31.7	56.3	1.7	15.1	27.6	41.3	48.0	15.2	9.5	18.3
MKL multi feature [31]	37.6	47.8	15.3	15.3	21.9	50.7	50.6	30.0	17.3	33.0
Best VOC07 [8]	26.2	40.9	9.8	9.4	21.4	39.3	43.2	24.0	12.8	14.0

	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	average
Our model	24.2	11.8	49.1	41.9	35.7	14.5	18.9	23.3	34.3	41.3	28.7
Our model (HOG only)	23.1	10.9	48.0	38.4	34.7	14.3	17.1	21.0	32.7	38.8	26.9
Our model (HOG only) no pre-filter	22.5	12.5	49.2	40.1	32.9	15.5	18.0	22.9	31.5	37.5	27.1
Our model (global)	10.4	5.8	35.0	32.0	20.1	12.1	13.5	11.3	24.1	29.7	19.6
DPM (VOC07) [10]	0.2	2.3	18.2	27.6	21.3	12.0	14.3	12.7	13.4	28.9	17.1
DPM [10]	22.8	10.6	44.1	37.0	35.2	13.6	16.1	18.5	31.8	36.9	25.9
Multi-class layout [5]	24.0	11.7	45.0	39.4	35.5	15.2	16.1	20.1	34.2	35.4	27.2
Bin. struct learning [26]	26.1	11.3	48.5	38.9	35.8	14.8	17.7	18.8	34.1	39.8	27.5
MKL multi feature [31]	22.5	21.5	51.2	45.5	23.3	12.4	23.9	28.5	45.3	48.5	32.1
Best VOC07 [8]	9.8	16.2	33.5	37.5	22.1	12.0	17.5	14.7	33.4	28.9	23.3

Table 2. Results on the PASCAL VOC 2007 object detection challenge.

point variation, and provides an implicit occlusion reasoning. Quantitatively our scheme achieves competitive performance on the difficult PASCAL VOC 2007 challenge, and qualitatively yields object segmentations with meaningful part labelings that reoccur across object instances. In future work we will investigate adding more complementary features and a notion of global context.

Acknowledgments This work has been funded, in part, by GRK 1362 of the German Research Foundation (DFG).

References

[1] Y. Amit and A. Trouve. POP: Patchwork of parts models for object recognition. *IJCV*, 75(2):267–282, 2007.

[2] G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. In *CVPR'05*.

[3] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV'06*.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.

[5] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV'09*.

[6] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. In *ECCV'08*.

[7] B. Epshtein and S. Ullman. Semantic hierarchies for recognizing objects and parts. In *CVPR'07*.

[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL VOC Challenge 2007.

[9] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[10] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR'08*.

[11] R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale invariant learning. In *CVPR'03*.

[12] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22(1):67–92, 1973.

[13] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 8, 2007.

[14] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV'09*.

[15] T. Joachims. *Making Large-Scale SVM Learning Practical*. Advances in Kernel Methods, 1999.

[16] A. Kapoor and J. Winn. Located hidden random fields: Learning discriminative parts for object detection. In *ECCV'06*.

[17] M. Kumar, A. Zisserman, and P. Torr. Efficient discriminative learning of parts-based models. In *ICCV'09*.

[18] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'06*.

[19] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*.

[20] Z. Lin, G. Hua, and L. Davis. Multiple instance feature for robust part-based object detection. In *CVPR'09*.

[21] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[22] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR'08*.

[23] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *PAMI*, 29(10):1848–1852, 2007.

[24] M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. In *CVPR'08*.

[25] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *IJCV*, 56(3):151–177, 2004.

[26] P. Schnitzspan, M. Fritz, S. Roth, and B. Schiele. Discriminative structure learning of hierarchical representations for object detection. In *CVPR'09*.

[27] P. Schnitzspan, M. Fritz, and B. Schiele. Hierarchical support vector random fields: Joint training to combine local and global features. In *ECCV'08*.

[28] M. Stark, M. Goesele, and B. Schiele. A shape-based object class model for knowledge transfer. In *ICCV'09*.

[29] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1–3):291–330, 2008.

[30] D. Tran and D. Forsyth. Configuration estimates improve pedestrian finding. In *NIPS'07*.

[31] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV'09*.