What Makes a Good Detector? – Structured Priors for Learning from Few Examples

Tianshi Gao[1], Michael Stark[1,2], and Daphne Koller[1]

[1] Stanford University
[2] Max Planck Institute for Informatics

Abstract. Transfer learning can counter the heavy-tailed nature of the distribution of training examples over object classes. Here, we study transfer learning for object class detection. Starting from the intuition that "what makes a good detector" should manifest itself in the form of repeatable statistics over existing "good" detectors, we design a low-level feature model that can be used as a prior for learning new object class models from scarce training data. Our priors are structured, capturing dependencies both on the level of individual features and spatially neighboring pairs of features. We confirm experimentally the connection between the information captured by our priors and "good" detectors as well as the connection to transfer learning from sources of different quality. We give an in-depth analysis of our priors on a subset of the challenging PASCAL VOC 2007 data set and demonstrate improved average performance over all 20 classes, achieved without manual intervention.

1 Introduction

Object class recognition has achieved remarkable performance for a wide variety of object classes [1]. The simultaneous recognition of many classes remains a challenging problem, however, due to both increasing model complexity and the required amount of training data. While image data is abundant for some classes, recent studies [2–4] confirm the heavy-tailed nature of their distribution over categories, fueling the need for learning algorithms that make efficient use of scarce training data. For image-level object classification, transfer learning has been widely adopted as a promising route towards reducing the amount of required training images, by re-using knowledge from existing object class models in the learning of new models. Transfer learning approaches differ in the particular representation of transferable knowledge, ranging from shared features [5, 6] over visual attributes [7–10] to classifiers at different levels in a hierarchy [11–13]. They typically build on global image representations that are consistent across different object classes, facilitating transfer learning.

This is different for object class detection. Today's most successful detectors are based on either fully rigid [14] or deformable [15] templates that capture the precise spatial layout of visual features of the object class of interest. Their performance depends crucially on the proper alignment of training images. Transferring knowledge between these models thus faces the additional challenge of

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 354–367, 2012.

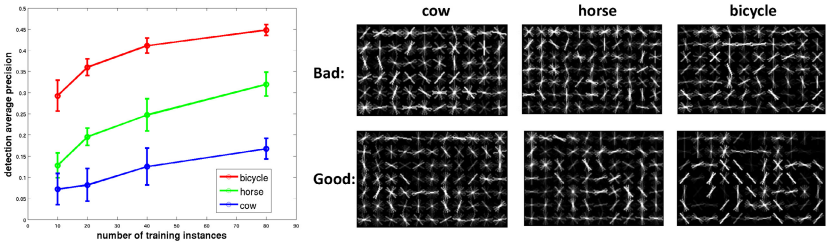

Fig. 1. Left: Detection results on VOC2007 test set with varying number of training instances. Right: Visualization of the learned models with low AP versus high AP.

aligning models prior to performing the transfer, i.e., establishing correspondences between source and target models. While there have been attempts to establish these correspondences by restricting transfer to topologically similar classes and views [16], annotating parts [17], or borrowing or sharing those training examples that happen to be well aligned [3, 4], the question "which knowledge to transfer where" has not been answered in a principled fashion.

In this paper, we therefore refrain from global transfer learning requiring model correspondences, and instead focus on local knowledge on the level of individual features or small local feature arrangements. This local knowledge is generic enough to be applicable in a spatially invariant fashion (as, e.g., pairwise potentials in an MRF for segmentation), without the need for correspondences.

The starting-point of our approach is the intuition that "what makes a good detector" – its ability to emphasize features that appear consistently across many training examples of an object class while suppressing background noise – should be evident (and measurable) for a given detector. In particular, we observe non-accidental structures to manifest increasingly as the number of training images and hence detector quality is increased (see Fig. 1). Those structures comprise very local details such as dominant gradient orientations as well as slightly bigger spatial structures, such as neighboring gradients forming continued line segments, corners, or parallel structures. Based on this intuition, we design a local correlation model that aims to capture these non-accidental structures that seem indicative of "good" detectors. It can be learned from one or several "good" existing detectors, and used as an informative prior to facilitate the learning of new detectors from few training examples.

Our paper makes the following contributions. First, we approach the difficult problem of transfer learning for object class detection by designing local-level spatial priors that circumvent the need for establishing across-model correspondences, based on today's de-facto standard detector, the DPM [15]. Second, we show that our priors capture structural information that is indicative of "good detectors", and can hence be used as measuring devices in order to assess the performance of a given object class detector. For transfer learning, our priors can predict the success of transferring knowledge from a specific source to a specific target. And third, we perform an in-depth experimental study that demonstrates

the ability of our priors to boost the performance of new detectors learned from few training examples without manual intervention.

2 Related Work

Transfer learning has received increasing attention in the recent literature, mainly following three different directions.

A first direction assembles new object class models from (components of) existing ones. Bart and Ullman [18] replace individual features of existing models in order to learn new classes, while Torralba et al. [5] learn a shared feature representation jointly from all classes. An entire branch of literature is dedicated to forming object class models from attributes using generative (Ferrari and Zisserman [19], Lampert et al. [7]) as well as discriminative (Wang and Mori [9], Farhadi et al. [8]) techniques. Others determine viable attribute combinations from linguistic knowledge bases (Rohrbach et al. [10]) and the web (Berg et al. [20]). On the highest level of abstraction, entire object class models are combined using stacking (Tommasi et al. [21]), multiple kernel learning (Luo et al. [6]) or along paths of a hierarchy (Levi et al. [11], Zweig and Weinshall [12], Marszalek and Schmid [22], Li et al. [13], and Salakhutdinov et al. [3]).

A second direction represents new classes relative to a set of known classes. Bart and Ullman [23] characterize new classes by means of their distance to known classes in feature space. Wang et al. [2] reason about similarity differences between sets of classes. In the context of template-based object class detection, Aytar and Zisserman [16] enforce closeness between a newly learned and an existing model by means of ℓ_2 regularization. While their approach delivers remarkable performance for specific sets of classes, it is limited by the required strict global alignment of templates.

A third direction explicitly aims to capture and transfer variations observed in the training data between different classes on different levels of abstraction. Miller et al. [24] transfer reoccurring spatial transformations between characters for improved character recognition. Stark et al. [17] transfer variations in local shape and global geometric layout of a part-based model, based on manually established part correspondences. Fei-Fei et al. [25] consider a Bayesian framework for image classification, where the posterior distribution over entire object class models is adapted in response to incoming training data of a new class.

While our approach clearly follows the third direction, it is among few that consider transfer learning for the challenging detection task based on state-of-the-art detectors. To our knowledge, we are the first to consider spatial priors that can capture correlation structures between features for this task, going beyond the i.i.d. assumption of previous work [3, 16]. Further, in contrast to previous work relying on global template alignment [3, 16], our focus is on local information, which allows to leverage transfer even in adversary and generic settings in an interpretable way (see Sect. 6). In a Bayesian sense, these previous works transfer mean parameters with diagonal covariance, while we build more informative non-diagonal covariance – our contribution is orthogonal, and could

be combined with existing techniques to the mutual benefit of both. Constructing informative covariance for transfer learning has also been proposed by Raina et al. [26] for text classification and Elidan et al. [27] for object shape modeling.

3 Preliminaries

We build upon the state-of-the-art sliding window object detector [15] that represents an object class as a mixture of multiple components, and each component consists of a designated root and multiple part templates. We stress that the ability of our approach to handle the full-fledged part-based incarnation of this model is in contrast to recent prior work [16], where parts pose fundamental challenges due to required template alignment. For the clarity of presentation, we first restrict ourselves to single component mixtures without parts, but relax that assumption in Sect. 5 and all experiments.

An object template according to [15] is specified by the model parameter $\mathbf{w}$. An image subwindow is represented by its feature vector $\mathbf{x}$ and scored by the inner product $\mathbf{w}^T\mathbf{x}$. More specifically, given an image subwindow, it is first spatially divided to $m \times n$ cells, and a histogram of oriented gradient is computed for each cell [14]. Thus, the feature vector consists of a collection of cell-level features, i.e., $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_{mn}] \in \mathbb{R}^{mnl}$, where $\forall i \in \{1, \ldots, mn\}$, $\mathbf{x}_i \in \mathbb{R}^l$ and l is the dimension of the cell-level feature. We use ";" to denote the vertical concatenation of vectors. Similarly, the model parameter can be decomposed in the same way $\mathbf{w} = [\mathbf{w}_1; \mathbf{w}_2; \ldots; \mathbf{w}_{mn}] \in \mathbb{R}^{mnl}$. We call each $\mathbf{w}_i$ a *cell model*. The visualization of an object model is shown in Fig. 2 (a). We can learn the object template by collecting a set of labeled training instances $(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$, and minimizing the following loss function:

$$\mathcal{L}(\mathbf{w}, \mathcal{X}, \mathcal{Y}) = \underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\mathcal{L}_{\text{prior}}(\mathbf{w})} + \underbrace{\frac{C}{N}\sum_{i=1}^{N} \max\{0, 1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)}\}}_{\mathcal{L}_{\text{data}}(\mathbf{w}, \mathcal{X}, \mathcal{Y})} \tag{1}$$

The loss function consists of two terms. The first is a hinge loss term $\mathcal{L}_{\text{data}}$ encouraging the model to classify training examples well. In a Bayesian interpretation, the second term $\mathcal{L}_{\text{prior}} = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ encodes the prior belief that the model is drawn from a zero-mean isotropic Gaussian. This constitutes an unstructured prior, since it assumes that every coordinate in $\mathbf{w}$ is independent. The quality of the model learned by minimizing (1) depends heavily on the number of (positive) training instances (see Fig. 1). As can be seen, the test performance is low for the case of few training examples due to overfitting.

4 Constructing Informative Structured Priors

As motivated in the introduction, the starting-point of our approach is the attempt to capture local model statistics that are characteristic of "good" detectors

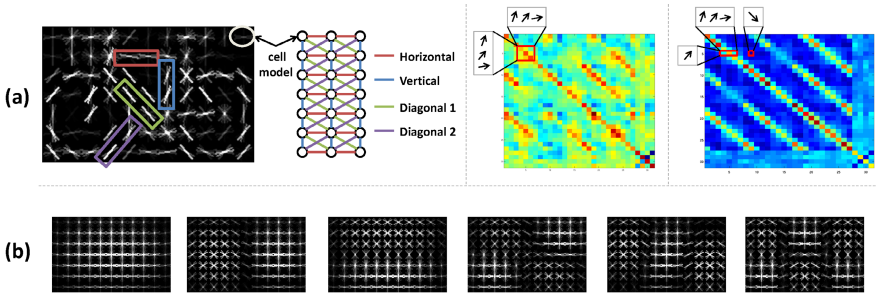

Fig. 2. (a) Left: Pairwise neighborhood. Middle: Cross covariance between horizontal pairs of cells. Right: Cell model covariance. **(b)** Eigen vectors of K_s (learned from horse) with increasing eigen values.

in the form of a prior. We hope (and verify in Sect. 6.2 and 6.3) that such a prior, when learned from an appropriate set of existing "good" detectors, manages to transfer these "good" characteristics to a target class of interest, for which we only have few training images available.

4.1 Local Correlation Structures in "Good" Object Class Detectors

According to Fig. 1 (right), we focus our attention on two distinct characteristics that we observe more prominently in the visualization of the "good" detectors than in the visualizations of poorer detectors.

First, we focus on examining individual cell models. We observe that activations of different gradient orientation bins do not seem entirely random, but rather correlated. Neighboring gradient orientation bins are often active together, while the majority is entirely suppressed. We attribute this observation to the fact that the template has to account for small variations in the local gradient directions in order to be robust. In addition, if a certain gradient orientation is encouraged, its orthogonal counterpart is often penalized. These are verified by the learned covariance matrix from good bicycle models, as shown in Fig. 2 (a)(right). It can be seen that similarly oriented gradients are usually positively correlated, while orthogonal gradients are often negatively correlated.

Second, we extend our focus to local neighborhoods of cells. Again, we observe obvious cross correlations between activations of gradient orientation bins. As shown in Fig. 2 (a)(middle), the cross covariance matrix has strong positive diagonal and correlations between nearby orientations. Dominant orientation bins of neighboring cells often follow similar patterns. They tend to coincide (forming line segments), to disagree by an angle (forming curves and corners), or to be roughly parallel, which we attribute to two causes. The first one is that the template has to be robust to small spatial variations in the alignment of training instances, causing neighboring cell models to show similar patterns. The second one is that the gradient-based nature of the underlying HOG features tends to capture object outlines, which are often smooth lines and curves.

We conclude that these *local correlation structures* on the level of both individual cells, spanning different gradient orientation bins, and on the level of neighboring cells, constitute valuable information that we would like to capture in our priors for use in transfer learning. This is in contrast to many off-the-shelf learning algorithms like SVM that represent the model as a flat vector, without considering the underlying structure.

4.2 Learning Structured Priors

We focus on transfer learning at the model parameter level. Specifically, suppose we want to learn a horse detector from a training set $(\mathcal{X}, \mathcal{Y})$ (the *target domain*). We are also given another set of labeled examples from some other categories $(\mathcal{X}_s, \mathcal{Y}_s)$, e.g., cow (the *source domain*). Then the goal is to construct an informative prior $\mathcal{L}_{\text{prior}}(\mathbf{w}, \mathcal{X}_s, \mathcal{Y}_s)$ for the target domain from the source. An obvious route taken by previous work [16] is to use the model $\mathbf{w}_s$ learned from the source as the mean for the target. In this case, the prior becomes $\mathcal{L}_{\text{prior}}(\mathbf{w}, \mathcal{X}_s, \mathcal{Y}_s) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}_s\|_2^2$, which assumes the strict global correspondances between the source and target. Moreover, it assumes the coordinates of $\mathbf{w}$ to be independent, similar to the uninformative prior $\frac{1}{2}\|\mathbf{w}\|_2^2$. In contrast, we aim to construct an informative *structured prior* $\mathcal{L}_{\text{prior}}^{\text{struct}}(\mathbf{w}, \mathcal{X}_s, \mathcal{Y}_s) = \frac{1}{2}\mathbf{w}^T K_s \mathbf{w}$ without correspondance assumptions.

We propose to learn both the within cell correlations, i.e., the cell model covariance, and pairwise correlations between neighboring cells, known as cross-covariance (we distinguish four different spatial relations according to Fig. 2 (a)).

We use a bootstrap technique to estimate both. Specifically, given the source domain $(\mathcal{X}_s, \mathcal{Y}_s)$, we generate multiple samples by randomly sampling S subsets of instances, resulting in $\{(\mathcal{X}_s^{(i)}, \mathcal{Y}_s^{(i)})\}_{i=1}^S$. For each $(\mathcal{X}_s^{(i)}, \mathcal{Y}_s^{(i)})$, we learn $\mathbf{w}_s^{(i)} = \text{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{X}_s^{(i)}, \mathcal{Y}_s^{(i)})$ as defined in (1). Each $\mathbf{w}_s^{(i)}$ gives us a set of cell-level pairwise neighbors $\mathcal{P}_t^{(i)} = \{(\mathbf{w}_{s,j}^{(i)}, \mathbf{w}_{s,k}^{(i)}) \mid (j,k) \in \mathcal{N}_t\}$ for each type of neighborhood relation $t \in \mathcal{T} = \{\text{horizontal}, \text{vertical}, \text{diag1}, \text{diag2}\}$. So the sample set we use to estimate the cross-covariance $\Sigma_t \in \mathbb{R}^{l \times l}$ is $\mathcal{P}_t = \cup_{i=1}^S \mathcal{P}_t^{(i)}$. Σ_t is then estimated by its sample average:

$$\Sigma_t = \frac{1}{|\mathcal{P}_t|} \sum_{(\mathbf{w}_{s,j}, \mathbf{w}_{s,k}) \in \mathcal{P}_t} (\mathbf{w}_{s,j} - \bar{\mathbf{w}}_{s,j})(\mathbf{w}_{s,k} - \bar{\mathbf{w}}_{s,k})^T \tag{2}$$

where $(\bar{\mathbf{w}}_{s,j}, \bar{\mathbf{w}}_{s,k})$ is the mean of all $(\mathbf{w}_{s,j}, \mathbf{w}_{s,k}) \in \mathcal{P}_t$. Similarly, we estimate the cell model covariance matrix $\Sigma_c \in \mathbb{R}^{l \times l}$ from its samples $\mathcal{P}_c = \cup_{i=1}^S \mathcal{P}_c^{(i)}$, where $\mathcal{P}_c^{(i)} = \{\mathbf{w}_{s,j}^{(i)} \mid \forall j \text{ indexing the cell model in } \mathbf{w}_s^{(i)}\}$.

Given the pairwise cross-covariance matrices Σ_t's, and the cell model covariance Σ_c, we construct the "covariance matrix" $\Sigma_s \in \mathbb{R}^{mnl \times mnl}$ for the target model, where m and n are the numbers of rows and columns in the target template. Σ_s is a $m \times n$ block matrix, where each block $\Sigma_{s,(j,k)} \in \mathbb{R}^{l \times l}$ is the cross-covariance between the j-th and the k-th cell models. Given a pair of cells (j,k) (assume $j \leq k$), $\Sigma_{s,(j,k)}$ can take on one of 6 different values depending

on the spatial relations between j and k. First, if $(j,k) \in \mathcal{N}_t$ for some $t \in \mathcal{T}$, then $\Sigma_{s,(j,k)} = \Sigma_t$ and $\Sigma_{s,(k,j)} = \Sigma_t^T$. Second, if $j = k$, then $\Sigma_{s,(j,k)} = \Sigma_c$, i.e., the covariance for cell models. Finally, if cell j and k are not spatial neighbors, $\Sigma_{s,(j,k)} = \Sigma_{s,(k,j)} = \mathbf{0}$, which implies that Σ_s is a *block sparse matrix*.

With Σ_s, we define our structured prior loss function as

$$\mathcal{L}_{\text{prior}}^{\text{struct}}(\mathbf{w}, \mathcal{X}_s, \mathcal{Y}_s) = \mathbf{w}^T(I - \lambda\Sigma_s)\mathbf{w} = \mathbf{w}^T K_s \mathbf{w} \tag{3}$$

where I is the identity matrix and $K_s = (I - \lambda\Sigma_s)$ [1]. λ is a scalar constant and serves two roles. First, it ensures a strongly convex prior, i.e., $K_s > 0$. λ scales the eigen values γ_i's of Σ_s such that all $1 - \lambda\gamma_i > 0$. Second, λ controls the degree of transfer. The larger λ, the stronger we transfer. For $\lambda = 0$, we get back to the uninformative prior $\mathbf{w}^T\mathbf{w}$. Note that K_s is still a *block sparse matrix*, where non-zero blocks correspond to those cell pairs that are spatial neighbors.

4.3 Interpretation

We interpret our prior (3) from different perspectives. First, we directly make sense of the informative part $\mathbf{w}^T\Sigma_s\mathbf{w}$. It can be decomposed as $\sum_{j=1}^{mn} \mathbf{w}_j^T \Sigma_c \mathbf{w}_j + \sum_{t\in\mathcal{T}} \sum_{(j,k)\in\mathcal{N}_t \text{ and } j<k} 2\mathbf{w}_j^T \Sigma_t \mathbf{w}_k$. There are two types of terms. The first type $\mathbf{w}_j^T \Sigma_c \mathbf{w}_j$ for each cell encourages the target cell model to follow the most likely directions (principal components) of the source cell models. The second type $\mathbf{w}_j^T \Sigma_t \mathbf{w}_k$ encourages a pair of cell models to follow the pairwise correlations estimated from the source. For example, if $\Sigma_{t,(q,r)} > 0$ (meaning that the q-th coordinate of one cell model is positively correlated with the r-th coordinate of its neighbor), we encourage the target $\mathbf{w}_{j,q}$ and $\mathbf{w}_{k,r}$ to take the same sign. Note that if two similar gradient orientations are often either both encouraged or both penalized for neighboring cells, this kind of "smoothing knowledge" is encoded in the prior and enforced for the target.

From a regularization point of view, consider the eigen decomposition $K_s = Q\Lambda Q^T$. Then we have $\mathbf{w}^T K_s \mathbf{w} = \|\Lambda^{\frac{1}{2}} Q^T \mathbf{w}\|_2^2$. Under the orthogonal transformation Q^T, instead of regularizing uniformly across all eigen directions $\|Q^T\mathbf{w}\|_2^2 = \|\mathbf{w}\|_2^2$, we penalize different directions of the eigen vectors proportional to their corresponding eigen values. Fig. 2 (b) shows the sorted eigen vectors of K_s learned from the horse model with increasing eigen values. As can be seen, these eigen vectors encourage spatially neighboring cell models to have similar patterns, which is the "smoothing effect" we expected.

From a probabilistic model point of view, the prior $\mathbf{w}^T K_s \mathbf{w}$ corresponds to a Gaussian Markov Random Field (GMRF) [28] with the graphical structure

[1] One tempting way to construct the prior is $\mathbf{w}^T \Sigma_s^{-1} \mathbf{w}$, assuming Σ_s is a valid covariance matrix. However, Σ_s may not be positive definite. We think of Σ_s as an affinity matrix (i.e., stronger correlation means higher affinity), which we use to directly construct the precision matrix of a Gaussian. We thus inherit a block sparsity structure that implies the desired conditional independencies. Also note that the regularization $-\mathbf{w}^T\Sigma_s\mathbf{w}$ penalizes different eigen directions in the same order as $\mathbf{w}^T\Sigma_s^{-1}\mathbf{w}$.

the same as the neighborhood system shown in Fig. 2 (a). It is related to discrete MRFs for image segmentation. Instead of using an MRF to capture the smoothness prior for nearby segment labels, we use the GMRF to capture the correlation between nearby gradient orientations and neighboring cell models.

5 Learning Detectors with Structured Priors

Given the prior learned from the source, we combine it with the data loss term to form the informed loss function as follows:

$$\mathcal{L}^{\text{struct}}(\mathbf{w}, \mathcal{X}, \mathcal{Y}) = \frac{1}{2}\mathbf{w}^T K_s \mathbf{w} + \frac{C}{N}\sum_{i=1}^{N} \max\{0, 1 - y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)}\} \tag{4}$$

Since the prior term is strongly convex (K_s is positive definite) and the data term is also convex, the objective (4) is convex. In fact, one can transform (4) to an equivalent regular SVM problem. Consider the eigen decomposition $K_s = Q\Lambda Q^T$. Define $\tilde{\mathbf{w}} = \Lambda^{\frac{1}{2}}Q^T\mathbf{w}$, then we have

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{w}} = \mathbf{w}^T Q\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}Q^T\mathbf{w} = \mathbf{w}^T Q\Lambda Q^T\mathbf{w} = \mathbf{w}^T K_s\mathbf{w} \tag{5}$$

Similarly, define $\tilde{\mathbf{x}} = \Lambda^{-\frac{1}{2}}Q^T\mathbf{x}$, then we have

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{x}} = (\Lambda^{\frac{1}{2}}Q^T\mathbf{w})^T(\Lambda^{-\frac{1}{2}}Q^T\mathbf{x}) = \mathbf{w}^T(Q\Lambda^{\frac{1}{2}}\Lambda^{-\frac{1}{2}}Q^T)\mathbf{x} = \mathbf{w}^T\mathbf{x} \tag{6}$$

Given the transformed feature $\tilde{\mathbf{x}}$ and parameter $\tilde{\mathbf{w}}$, the original (4) becomes:

$$\mathcal{L}^{\text{struct}}(\tilde{\mathbf{w}}, \mathcal{X}, \mathcal{Y}) = \frac{1}{2}\tilde{\mathbf{w}}^T\tilde{\mathbf{w}} + \frac{C}{N}\sum_{i=1}^{N} \max\{0, 1 - y^{(i)}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}^{(i)}\} \tag{7}$$

Note that this is a regular SVM problem and can be solved by any existing highly optimized algorithm or software package. We can recover the optimal solution of (4) as $\mathbf{w}^* = Q\Lambda^{-\frac{1}{2}}\tilde{\mathbf{w}}^*$.

The former extends naturally to a part-based model with a mixture of multiple components [15]. In this case, the model parameters consist of multiple blocks, one for each pair of part and component, i.e., $\mathbf{w} = [\mathbf{w}^{(1,1)}; \ldots; \mathbf{w}^{(v,p)}]$, where p is the number of parts and v is the number of components. We can construct a prior $K_s^{(i,j)}$ for each of them, and the overall structured prior is a block diagonal matrix $K_s = \text{diag}(K_s^{(1,1)}, \ldots, K_s^{(1,p)}, K_s^{(2,1)}, \ldots, K_s^{(v,p)})$.

6 Experiments

In this section, we carefully analyze the performance of our proposed structured priors, in three different settings. First (Sect. 6.2), we come back to the question of "what makes a good detector", by analyzing the likelihood of sequences of detectors of varying quality (learned from varying numbers of training images)

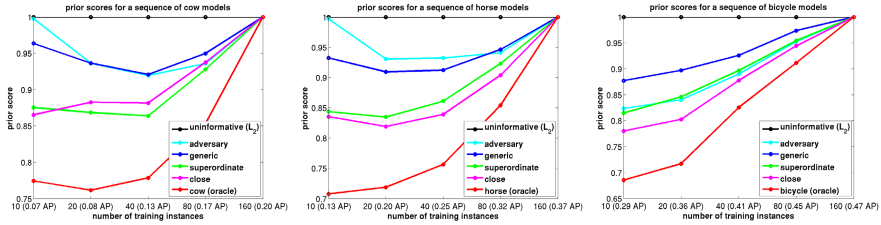

Fig. 3. Log likelihood scores for sequences of target models with increasing quality using both our priors learned from various sources and the uninformative prior

under our priors. We observe that model quality and model prior likelihood are in fact correlated. Second (Sect. 6.3), we apply our priors in a transfer learning task, where we learn target class models from few training examples plus priors learned from different sets of source classes. We give an in-depth analysis on a subset of PASCAL VOC 2007 classes and compare adversary, generic, superordinate, and close category priors. And third (Sect. 6.4), we switch to a real-world application scenario in which we demonstrate transfer learning for all 20 PASCAL classes without any human intervention, by using either generic priors or semantic relatedness measures mined from linguistic knowledge bases [10].

6.1 Basic Setup

We use the PASCAL VOC 2007 dataset [1] containing 20 object classes. Learning is done on the train/val sets and the entire test set is used for testing. We simulate scarce training data by randomly subsampling 5 times from the original training set, varying the number of training instances per component from 10, 20 to 40. The mean and standard deviation over 5 runs are reported.

The baseline is the latest implementation of the state-of-the-art detector [15]. We first invest CPU time into a deeper analysis of a simpler root-only variant in Sect. 6.2 and 6.3, and then perform a broader and extensive study on the fully part-based model for all categories in Sect. 6.4.Our priors are learned using bootstrap (see Sect. 4.2) with sample number 5 and sample set size 160 per component. The C parameter of SVM is set to the default (and optimized) 0.002 from [15] for the baseline as well as our models. We did not tune the λ parameter and set it such that the largest eigen value of $\lambda\Sigma_s$ is 0.9 (therefore, $K_s > 0$ in (3)) in all experiments.

6.2 Likelihood Analysis

Before applying our priors in an actual transfer learning task, we commence by giving an analysis of their log likelihood functions $-\frac{1}{2}\mathbf{w}^T K_s \mathbf{w}$ (ignoring the normalization constant). Specifically, we consider sequences of detectors of a target class, learned from varying numbers of training images (Fig. 3). Naturally, these detectors are of different quality – low numbers of training images result

in "poor", while high numbers result in "good" detectors (which seems obvious, and is in fact reflected by increasing AP values given as part of the x axis labels). We then evaluate the log likelihood of each detector under different variants of our priors. For each prior, we plot log likelihoods normalized by the highest attained log likelihood value in the sequences for comparison.

The purpose of this experiment is to highlight two key aspects of our approach. First, we want to verify that the specific form of structured priors that we are proposing does in fact capture relevant model structures that are indicative of "good" detectors (see Sect. 4.1), and are hence worthwhile considering for transfer. And second, we want to demonstrate that priors learned from different sets of source models exhibit different characteristics in their log likelihood scores w.r.t. a target class of interest, and can thus hint on the expected performance of newly learned models using the priors.

Results. As target classes, we choose ones that we deem representative for different levels of baseline performance (see Fig. 1), namely, *cow* (weak), *horse* (modest), and *bicycle* (strong). Fig. 3 plots the normalized log likelihood scores by applying various priors to sequences of detectors of the three target classes, learned from varying numbers of training examples (left: *cow*, middle: *horse*, right: *bicycle*). For all three target classes, we consider priors learned from sets of source classes of varying similarity to the respective target class, according to human judgment. We distinguish (1) *oracle priors* (red curves), learned from the target class itself, (2) *adversary priors* (cyan), learned from visually dissimilar classes among VOC2007 classes (*bus*), (3) *generic priors* (blue), learned from all but the target class, (4) *superordinate category priors* (green; *horse*, *sheep* for *cow*; *cow*, *sheep* for *horse*; *motorbike*, *bus*, *boat* for *bicycle*), learned from classes sharing the same superordinate class as the target class (omitting the target class itself), and (5) *close category priors* (magenta; *sheep* for *cow*; *cow* for *horse*; *motorbike* for *bicycle*) learned from the most similar class to the target class.

Indication of target detector quality. In Fig. 3, we first consider the oracle priors (red curves), learned from the respective target class of interest. While these priors are unrealistic in a transfer learning setting, we introduce them here for the sake of a best-case analysis. In Fig. 3, we observe that the log likelihood of detectors under the oracle priors (red curves) almost always increases with the number of training examples, and is hence positively correlated with the corresponding detection performance. This holds true for all three target classes (plots in Fig. 3 contain AP values as horizontal axis labels). I.e., the log likelihood values under our proposed prior do in fact provide an indication of the quality of the target class detector.

Indication of source quality. In Fig. 3, we further observe a fairly consistent ordering of log likelihood curves obtained by applying priors learned from different source classes to the respective target detectors. Specifically, the relative score gap between the best and the worse models and/or the degree of positive correlation between log likelihood values along a curve and the number of training examples increase from *uninformative* over *adversary*, *generic*, *superordinate*,

and *close category* to *oracle* priors, for all three target classes (Fig. 3 left, middle, and right), matching the intuitively expected ordering.

6.3 Transfer Learning

We continue by analyzing our structured priors in a transfer learning setting. To that end, we follow the selection of source and target classes of Sect. 6.2. Tab. 1 (a), 1 (b), and 1 (c) give the detection results (mean average precision and standard deviation) for target classes *cow*, *horse*, and *bicycle*, respectively. In Tab. 1 (a) (*cow*, weak baseline performance), we observe that the performance of the baseline (leftmost column) can be improved by all variants of our structured priors for all numbers of training images, with the only exception being *adversary priors* for 10 training examples. The largest improvements are achieved by *close category priors* (rightmost column) for 10 (3.93%) and by the *superordinate category priors* for 20 (3.49%) and 40 (2.40%) training examples. The *generic priors* consistently improve over the baseline (by 1.92% for 10, by 2.51% for 20, and by 1.83% for 40 training examples). Strikingly, even the adversary prior helps in two out of three cases (improving by 0.25% for 20 and 1.46% for 40 training examples).

A similar tendency can be observed in Tab. 1 (b) (*horse*, modest baseline performance). Our structured priors improve over the baseline for all numbers of training images. *Close category priors* improve by 1.5% for 10, *superordinate category priors* by 2.25% for 20, and *generic priors* by 0.72% for 40 training examples. Again, *generic priors* consistently improve over the baseline (by 0.82% for 10, by 1.74% for 20, and by 0.72% for 40 training examples). The *adversary* priors improve by 0.27% for 10 training examples, and are comparable to the baseline for 20 (−0.16%) and 40 (−0.15%) training examples.

For *bicycle* (Tab. 1 (c)), our structured priors can improve over the strong baseline for 10 and 20 training images in the *close category* incarnation (by 0.22% for 10 and 0.37% for 20 training images). Only for the case of 40 training images, the baseline detector already performs on a remarkable level of 41.09% AP, which we miss by 0.5% using *close category priors*. Because of the strong baseline, *generic* and *adversary category priors* can not further improve performance.

Connection to prior likelihood functions. For all three target classes (*cow*, *horse*, and *bicycle*), we observe a trend in the ordering of performance values across all numbers of training images. Performance typically increases from *adversary* over *generic* and *superordinate* to *close category* priors, which is consistent with the ordering established by means of prior log likelihood functions in Sect. 6.2, and hence can be used as a predictor for transfer learning success (at least in theory, since a sequence of target class models needs to be available for the prediction).

Summary. We conclude that transfer learning using our structured priors in fact improves performance over the baseline in many cases, in particular for classes with weak (*cow*) and modest (*horse*) baseline performance. Surprisingly, even the *generic priors* provide valuable information for learning these classes, leading to improved performance for all numbers of training images. *Close category priors*

Table 1. Detection results using models with various priors

		training #	baseline	adversary	generic	superordinate	close
(a)	cow	10	7.19 ± 3.69	6.49 ± 3.21	9.11 ± 3.09	10.29 ± 2.15	$\mathbf{11.12 \pm 2.42}$
		20	8.17 ± 3.88	8.42 ± 4.21	10.68 ± 3.85	$\mathbf{11.66 \pm 3.30}$	10.90 ± 3.13
		40	12.51 ± 4.36	13.97 ± 2.61	14.34 ± 3.22	$\mathbf{14.91 \pm 2.10}$	14.32 ± 2.84
(b)	horse	10	12.79 ± 2.94	13.06 ± 2.64	13.61 ± 1.03	13.37 ± 3.97	$\mathbf{14.29 \pm 2.80}$
		20	19.55 ± 2.06	19.39 ± 1.93	21.29 ± 1.91	$\mathbf{21.80 \pm 2.21}$	21.72 ± 1.97
		40	24.70 ± 3.82	24.55 ± 3.47	$\mathbf{25.42 \pm 2.80}$	23.76 ± 4.26	24.01 ± 3.44
(c)	bicycle	10	29.25 ± 3.67	27.64 ± 6.47	27.72 ± 5.68	28.60 ± 5.39	$\mathbf{29.47 \pm 5.30}$
		20	35.98 ± 2.01	35.11 ± 2.96	34.95 ± 2.79	35.87 ± 2.05	$\mathbf{36.35 \pm 2.42}$
		40	$\mathbf{41.09 \pm 1.80}$	40.31 ± 2.36	40.33 ± 1.23	40.26 ± 2.35	40.60 ± 2.03

Table 2. Results on all 20 categories using uninformative, generic and semantic priors

#		aero	cat	pers	bike	chair	plant	bird	cow	sheep	boat	table	sofa	bott	dog	train	bus	horse	tv	car	mbike	avg
10	base	**18.9**	7.1	0.9	39.9	0.7	1.2	0.9	11.8	6.0	5.4	14.4	12.6	8.4	1.0	20.2	20.3	14.7	27.9	**5.3**	**22.5**	12.0
	gen.	16.3	11.2	0.3	39.5	1.7	6.5	**4.0**	**18.2**	**18.0**	11.2	**18.4**	13.5	9.1	9.1	26.8	**47.0**	**32.8**	32.6	4.3	21.0	17.1
	sem.	16.4	**12.7**	**2.1**	**47.5**	**5.1**	**10.8**	2.2	17.9	14.7	**12.3**	17.0	**15.2**	**11.8**	**11.3**	**28.4**	46.1	32.0	**32.6**	4.9	21.6	**18.1**
20	base	19.8	10.5	**1.5**	39.7	5.9	7.9	**7.2**	18.5	13.9	13.3	14.8	17.3	15.7	6.5	**37.4**	42.0	42.3	37.4	11.6	37.7	20.0
	gen.	21.8	**15.4**	0.6	53.3	12.4	**9.6**	4.8	**23.4**	19.6	**14.3**	**18.9**	24.7	20.7	9.4	24.5	48.1	36.9	37.5	21.8	**38.4**	22.8
	sem.	**23.4**	14.1	0.7	**53.5**	**13.6**	7.9	6.8	22.6	**20.9**	14.3	17.7	**25.3**	**21.1**	**9.8**	26.9	**48.5**	**43.3**	**37.6**	**23.7**	34.1	**23.3**
40	base	25.7	14.4	4.1	53.4	11.9	10.7	6.4	20.4	17.7	**14.2**	21.6	21.8	16.6	**8.8**	**38.1**	46.4	46.7	**38.9**	24.2	40.6	24.1
	gen.	**26.9**	**16.5**	4.0	**56.3**	**14.6**	**11.9**	**8.6**	23.9	**19.0**	13.7	**23.6**	20.4	**18.0**	6.5	23.7	47.3	**53.0**	37.2	35.2	**42.6**	**25.1**
	sem.	26.0	15.5	**6.1**	55.6	13.5	7.0	8.3	**24.0**	18.1	14.1	18.3	**22.5**	9.6	7.3	26.3	47.4	51.5	34.7	**44.8**	39.9	24.5

improve over the baseline for all object classes for 10 and 20 training images, and in 2 of 3 classes (*cow*, *horse*) also for 40 training images.

6.4 Transfer Learning Using Semantic Relatedness

Motivated by the good performance of both *close category* and *generic* priors, we now extend the evaluation from the prototypical set of object classes in the previous section to the full set of VOC 2007 classes. Furthermore, we now switch to the full part-based model [15], in contrast to prior work limited to small sets of classes only [16, 17] or to inferior root-template-only models [3, 16]. We aim at evaluating transfer learning using our structured priors in a realistic setting, not using any human intervention. To that end, we propose the use of *Semantic Relatedness* (SR) measures in order to determine which source classes to consider as *close category priors* for a given target class, thus determining sources and targets of knowledge transfer fully automatically. Specifically, we determine the SR between all pairs of the 20 Pascal classes by querying Wordnet [29] and Wikipedia [30], using the implementation of [10]. This yields two continuous-valued similarity matrices, which we average in order to increase robustness [10]. We determine the set of the 3 most semantically related classes to each target class, which we use as the basis for learning *close category priors*. To our knowledge, we are the first to consider SR for transfer learning in object class detection. In addition, we consider *generic priors* that do not require any other information than the set of all class labels, promoting fully automatic transfer.

Results. We give the results of both transfer learning using *close category priors* from SR and *generic priors* in Tab. 2 (mean over 5 runs), and make the following

observations. First, transfer learning improves the average performance over all 20 Pascal classes compared to the baseline for all numbers of training images. *Generic priors* dominate for 40 (improvement over baseline 1%), *close category priors* for 10 (6.1% improvement) and 20 training images (3.3% improvement). Put differently, transfer learning wins in 50 of the considered 60 combinations of object classes and numbers of training images. *close category priors* improve over the baseline for 17 (10 training images) and 15 (20 training images) classes, respectively. Furthermore, note that the improvement is more pronounced for the part-based model than for the root-only model (e.g., 6.4% vs. 1.9% for the cow class, *generic priors*, and 10 training images). Second, the intuition that more specific priors should help more is reflected by the ordering of recognition performance. *Close category priors* win for 26 and *generic priors* for 24 combinations. The set of available close categories is of course very limited for the 20 Pascal classes. We thus expect further improvements for larger pools of classes.

7 Conclusions

We have considered transfer learning for object class detection, based on structured priors that we showed to be indicative of detector quality. Using these priors, we could consistently improve over the performance of a baseline detector for different transfer learning scenarios, ranging from close category over superordinate category to generic and even adversary priors. Leveraging the generic nature of our priors, we could improve on average over all classes of the challenging PASCAL VOC 2007 data set in a fully automatic fashion.

Acknowledgment. This work was supported by the Max Planck Center for Visual Computing and Communication, the NSF under grant No. RI-0917151, the Office of Naval Research MURI grant N00014-10-10933, and the Boeing company. We thank Marcus Rohrbach for semantic relatedness computations.

References

1. Everingham, M., van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. In: IJCV (2010)
2. Wang, G., Forsyth, D., Hoiem, D.: Comparative object similarity for improved recognition with few or no examples. In: CVPR (2010)
3. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. In: CVPR (2011)
4. Lim, J.J., Salakhutdinov, R., Torralba, A.: Transfer learning by borrowing examples for multiclass object detection. In: NIPS (2011)
5. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. In: CVPR (2004)
6. Luo, J., Tommasi, T., Caputo, B.: Multiclass transfer learning from unconstrained priors. In: ICCV (2011)
7. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)

8. Farhadi, A., Endres, I., Hoiem, D.: Attribute-centric recognition for cross-category generalization. In: CVPR (2010)
9. Wang, Y., Mori, G.: A Discriminative Latent Model of Object Classes and Attributes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 155–168. Springer, Heidelberg (2010)
10. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR (2011)
11. Levi, K., Fink, M., Weiss, Y.: Learning from a small number of training examples by exploiting object categories. In: LCVPR (2004)
12. Zweig, A., Weinshall, D.: Exploiting object hierarchy: Combining models from different category levels. In: ICCV (2007)
13. Li, L.J., Wang, C., Lim, Y., Blei, D., Fei-Fei, L.: Building and using a semantivisual image hierarchy. In: CVPR (2010)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
15. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2010)
16. Aytar, Y., Zisserman, A.: Tabula rasa: Model transfer for object category detection. In: ICCV (2011)
17. Stark, M., Goesele, M., Schiele, B.: A shape-based object class model for knowledge transfer. In: ICCV (2009)
18. Bart, E., Ullman, S.: Cross-generalization: Learning novel classes from a single example by feature replacement. In: CVPR (2005)
19. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
20. Berg, T.L., Berg, A.C., Shih, J.: Automatic Attribute Discovery and Characterization from Noisy Web Data. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 663–676. Springer, Heidelberg (2010)
21. Tommasi, T., Orabona, F., Caputo, B.: Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In: CVPR (2010)
22. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR (2007)
23. Bart, E., Ullman, S.: Single-example learning of novel classes using representation by similarity. In: BMVC (2005)
24. Miller, E., Matsakis, N., Viola, P.: Learning from One Example Through Shared Densities on Transforms. In: CVPR (2000)
25. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. PAMI 28, 594–611 (2006)
26. Raina, R., Ng, A.Y., Koller, D.: Constructing informative priors using transfer learning. In: ICML (2006)
27. Elidan, G., Packer, B., Heitz, G., Koller, D.: Convex point estimation using undirected bayesian transfer hierarchies. In: UAI (2008)
28. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
29. Fellbaum, C.: WordNet: An Electronical Lexical Database. The MIT Press (1998)
30. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: IJCAI (2007)