On Taxonomies for Multi-class Image Categorization

**Alexander Binder · Klaus-Robert Müller ·
Motoaki Kawanabe**

Abstract We study the problem of classifying images into
a given, pre-determined taxonomy. This task can be ele-
gantly translated into the structured learning framework.
However, despite its power, structured learning has known
limits in scalability due to its high memory requirements
and slow training process. We propose an efficient approx-
imation of the structured learning approach by an ensem-
ble of local support vector machines (SVMs) that can be
trained efficiently with standard techniques. A first theoret-
ical discussion and experiments on toy-data allow to shed
light onto why taxonomy-based classification can outper-
form taxonomy-free approaches and why an appropriately
combined ensemble of local SVMs might be of high practi-
cal use. Further empirical results on subsets of Caltech256
and VOC2006 data indeed show that our local SVM for-
mulation can effectively exploit the taxonomy structure and
thus outperforms standard multi-class classification algo-
rithms while it achieves on par results with taxonomy-based
structured algorithms at a significantly decreased computing
time.

A. Binder (✉) · K.-R. Müller · M. Kawanabe
Dep. Computer Science, Machine Learning Group, Berlin
Institute of Technology, Franklinstr. 28/29, 10587 Berlin,
Germany
e-mail: alexander.binder@tu-berlin.de

K.-R. Müller
e-mail: klaus-robert.mueller@tu-berlin.de

A. Binder · M. Kawanabe
Dep. Intelligent Data Analysis, Fraunhofer FIRST, Kekuléstr. 7,
12489 Berlin, Germany

M. Kawanabe
e-mail: motoaki.kawanabe@first.fraunhofer.de

1 Introduction

In computer vision, one of the most difficult challenges is
to bridge the semantic gap between appearances of image
contents and high-level semantic concepts (Smeulders et al.
2000). While systems for image annotation and content-
based image retrieval are continuously progressing, they are
still far from resembling the recognition abilities of humans
that have closed this gap. Humans are known to exploit tax-
onomical hierarchies in order to recognize general seman-
tic contents accurately and efficiently. Therefore, it remains
important for artificial systems to incorporate extra sources
of information, such as user tags[1] (Barnard et al. 2003;
Qi et al. 2009) or prior knowledge such as taxonomical rela-
tions between visual concepts.

There have been a number of studies considering *learn-
ing* class-hierarchies, for instance on the basis of delayed
decisions (Marszalek and Schmid 2008), dependency graphs
and co-occurrences (Lampert and Blaschko 2008; Blaschko
and Gretton 2009), greedy margin-trees (Tibshirani and
Hastie 2007), by hierarchical clustering (Fan 2005; Griffin
and Perona 2008), and by incorporating additional infor-
mation (Marszalek and Schmid 2007). Unfortunately, few
could so far report significant performance gains in the final
object classification (even though they contributed to other
aspects, for instance, computational efficiency).

When a taxonomy is available, a standard way of using
the hierarchy is sequential greedy decision (Griffin and Per-
ona 2008). Starting from the root node, the strategy selects

---

[1]Flickr. http://www.flickr.com.

only the most probable child at each node and ignores other possibilities until reaching a leaf node. Therefore, for classifying an unseen image only the classifiers on one path of the hierarchy need to be evaluated. Furthermore, since each node takes only relevant images for current and future decisions during the training phase, such greedy methods are computationally very attractive. The work in Griffin and Perona (2008) focuses on learning hierarchies and demonstrates speed gains by the greedy classification schemes compared to one versus all classifiers (e.g. 5-fold speed gain at the cost of 10% performance drop). Another greedy walk approach over a learned hierarchy (Marszalek and Schmid 2008) shows small improvements on the Caltech256 dataset.

In this work, we contribute a tractable alternative to the structure learning framework which can solve our task in a sophisticated way, but is less time consuming. We propose its efficient decomposition into an ensemble of local support vector machines (SVMs) that can be trained efficiently. Since the primal goal of this paper is to discuss how much and why pre-determined taxonomies improve classification performance, we consider any techniques for speed-up which degrade performance to be out of the scope of this paper.[2]

Our work is similar in spirit to Zweig and Weinshall (2007) who deployed user-determined taxonomies and showed that classifiers for super-classes at parents and grand-parents nodes can enhance leaf-node classifiers by controlling the bias-variance trade-off. However in Zweig and Weinshall (2007) the discrimination of images was performed against a small set of common backgrounds, and thus, all upper-node classifiers share the same negative samples, i.e. the background images. Performance was measured for object versus background scenarios. In contrast to Zweig and Weinshall (2007), we will study a more difficult problem, namely, multi-task or multi-label classification between object categories. Since our problem does not contain uniform sets of background, it is an interesting question whether an averaging along the leaves of a taxonomy integrating everything from super-class classifiers until the lower leaf-nodes can still help to improve the object recognition result, in particular as the negative samples can not be shared among all classifiers as in Zweig and Weinshall (2007).

We remark furthermore that we observe from our experiments that greedy strategies as e.g. Griffin and Perona (2008) are inferior to our novel taxonomy based methods that we propose in this paper.

---

[2]For instance, we use all images for SVM training at every node, which is of course more costly than the greedy strategy. It may be possible reducing the large number of negative examples which are inferred irrelevant to current and future decisions with high probability without decreasing classification accuracy.

In contrast to this work the above mentioned approaches have one aspect common in their methodology: they restrict performance measurement to flat loss measures which do not distinguish between different types of misclassification. In contrast to that humans tend to perceive some confusions like cat versus fridge to be more unnatural than others like cat versus dog which can be reflected by a taxonomy. The hierarchy in Griffin and Perona (2008) *learned* from features reflects feature similarities and is as a consequence in part not biologically plausible: the gorilla is closer to a raccoon than to a chimpanzee, the grasshopper is closest to penguin, and more distant to other insect lifeforms. Such problems can arise generally when the hierarchy is learned from image contents.

This prompts the question whether it is useful to employ a taxonomy which is based merely on information already present in the images and which is thus implicitly already in use through the extracted feature sets that feed the learning machine. Furthermore basic information derived from the images only, may not always be coherent with the user's rich body of experience and implicit or explicit knowledge.

An example is the discrimination of several Protostomia, sea cucumbers and fish (see Fig. 1). While sea cucumbers look definitely more similar to many Protostomia, they are much closer to fish sharing the property of belonging to Deuterostomia according to phylogenetic systematics. Equally, horseshoe crabs look more similar to crabs as both have a shell and live on the coast, but the horseshoe-crab as a member of Chelicerata is closer to spiders than to crabs. Therefore, this work is focused on *pre-determined* taxonomies constructed independently from basic image features as a way for providing such additional information rsp. knowledge. This task fits well into the popular structured learning framework (Taskar et al. 2004; Tsochantaridis et al. 2005) which has recently seen many applications among them in particular document classification with taxonomies (Cai and Hofmann 2004). Note furthermore that a given taxonomy permits to deduce a *taxonomy* loss function which—in contrast to the common 0/1 loss—allows to weight misclassification unevenly according to their mismatch when measured in the taxonomy. Thus, it is rather natural to evaluate classification results according to the taxonomy losses instead of the flat 0/1 loss, in this sense imposing a more human-like error measure.

The remainder of this paper is organized as follows. In Sect. 2 we will explain our novel local procedures with scoring deduced from generalized *p*-means, along with structure learning approaches. We discuss in Sect. 3 when and why our procedures can improve the one-vs-all baseline. The empirical comparisons between our local approach and other taxonomical algorithms and taxonomy-free baselines are presented in Sect. 4. For the present work, we have constructed multi-class classification datasets with taxonomy

Fig. 1 Mismatch between taxonomy and visual similarity: the first column are Protostomia, the second (sea cucumbers) and third row are Deuterostomia. The difference is based on embryonal development

trees between object categories based on the benchmarks Caltech256 (Griffin et al. 2007) and VOC2006 (Everingham et al. 2006) as explained in Sect. 4.1. In this section we discuss why our local approach can improve the one-vs-all baseline from the viewpoint of averaging processes. Section 6 gives concluding remarks and a discussion.

2 Learning Machines with Taxonomies for Multi-class Categorization

2.1 Problem Formulation

We consider the following problem setting: given are n pairs $\{(x^{(i)}, y^{(i)})\}$, $1 \leq i \leq n$, where $x^{(i)} \in \Re^d$ denotes the vectorial representation of the i-th image which can be represented in higher dimensions by a possibly non-linear mapping $\phi(x^{(i)})$. The latter gives also rise to a kernel function on images, given by $K_X(x, x') = \langle \phi(x), \phi(x') \rangle$. The set of labels is denoted by $Y = \{c_1, c_2, \ldots, c_k\}$. We focus on multi-class classification tasks, where every image is annotated by exactly one element of Y.[3]

In addition, we are given a taxonomy T in form of an arbitrary directed graph (V, E) where $V = (v_1, \ldots, v_{|V|})$ and $Y \subset V$ such that classes are identified with leaf nodes (see Fig. 2 for an example). We assume the existence of one

---

[3] Some image databases fall into the multi-label setting, where an image can be annotated with several class labels.

unique root node. The set of nodes on the path from the root node to a leaf node y is defined as $\pi(y)$. Alternatively, the set $\pi(y)$ can be represented by a vector $\kappa(y)$ where the j-th element is given by

$$\kappa_j(y) = \begin{cases} 1 & v_j \in \pi(y), \\ 0 & \text{otherwise,} \end{cases}$$

such that the category *sheep* in Fig. 2 is represented by the vector

$$\kappa(\text{sheep}) = (1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0,$$
$$0, 0)'.$$

The goal is to find a function f that minimizes the generalization error $R(f)$,

$$R(f) = \int_{\Re^d \times Y} \delta(y, f(x)) dP(x, y),$$

where $P(x, y)$ is the (unknown) distribution of images and annotations. The quality of f is measured by an appropriate, symmetric, non-negative loss function $\delta : Y \times Y \to \Re_0^+$ detailing the distance between the true class y and the prediction. For instance, δ may be the common 0/1 loss, given by

$$\delta_{0/1}(y, \hat{y}) = \begin{cases} 0 & y = \hat{y}, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

Fig. 2 Taxonomy constructed from VOC2006 labels. The life subtree is based on biological systematics
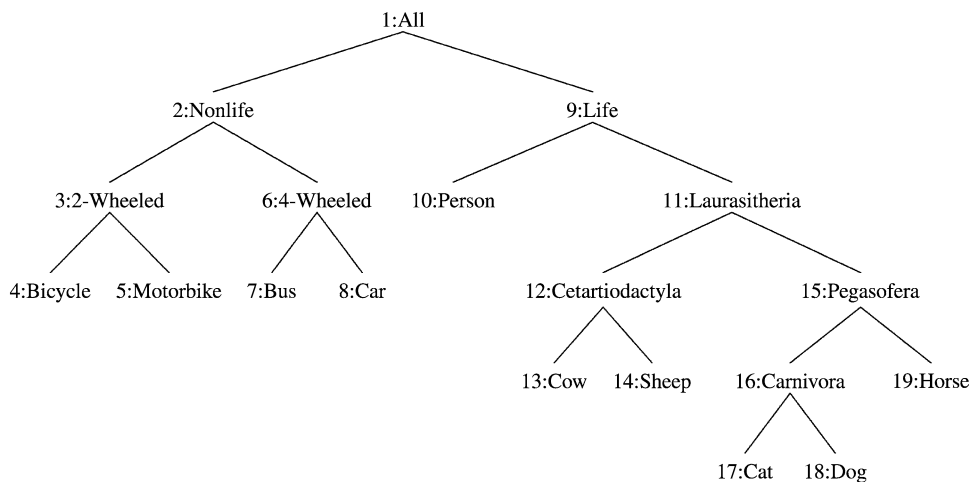
When learning with taxonomies, the distance of y and $\hat{y}$ with respect to the taxonomy is fundamental. For instance, confusing an *bus* with a *cat* is more severe than confusing the classes *cat* and *dog*. We will therefore also utilize a taxonomy-based loss function reflecting this intuition by counting the number of non-shared nodes on the path between the true class y and the prediction $\hat{y}$,

$$\delta_T(y, \hat{y}) = \sum_{j=1}^{|V|} |\kappa_j(y) - \kappa_j(\hat{y})|. \tag{2}$$

This distance can be induced as Hilbert space norm by the kernel between labels defined as

$$K_Y(y, \hat{y}) = \sum_{j=1}^{|V|} \kappa_j(y)\kappa_j(\hat{y}). \tag{3}$$

For instance, the taxonomy-based loss between categories *horse* and *cow* in Fig. 2 is $\delta_T(horse, cow) = 4$ because $\kappa(horse)$ and $\kappa(cow)$ differ at the nodes horse, pegasofera, cetartiodactyla and cow.

2.2 Structure Learning with Taxonomies

The taxonomy-based learning task can be framed as structured learning problem (Taskar et al. 2004; Tsochantaridis et al. 2005) where a function

$$f(x) = \arg\max_y \langle w, \Psi(x, y) \rangle \tag{4}$$

defined jointly on inputs and outputs is to be learned. The mapping $\Psi(x, y)$ is often called the joint feature representation and for learning taxonomies given by the tensor product (Cai and Hofmann 2004) with indicator functions

$[[v_i \in \pi(y)]]$

$$\Psi(x, y) = \phi(x) \otimes \kappa(y) = \begin{pmatrix} \phi(x)[[v_1 \in \pi(y)]] \\ \phi(x)[[v_2 \in \pi(y)]] \\ \vdots \\ \phi(x)[[v_{|V|} \in \pi(y)]] \end{pmatrix}.$$

Thus, the joint feature representation subsumes the structural information and explicitly encodes paths in the taxonomy. It leads to a joint kernel

$$K_{X,Y}((x_1, y_1), (x_2, y_2)) = K_X(x_1, x_2)K_Y(y_1, y_2), \tag{5}$$

where $K_X(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$ and the label kernel $K_Y(y_1, y_2)$ is defined according to the taxonomy T as in (3).

The empirical risk can be optimized utilizing conditional random fields (CRFs) (Lafferty et al. 2004) or structural support vector machines (SVMs) (Taskar et al. 2004; Tsochantaridis et al. 2005). We will follow structural learning in the formulation by Weston and Watkins (1999), Har-Peled et al. (2002). There are two ways of incorporating a loss $\Delta(y, \bar{y})$ such as $\delta_{0/1}$ and δ_T in the structural SVMs. The optimization problem with margin rescaling is given by

$$\min_{w,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi^{(i)}$$

$$\text{s.t.} \quad \forall i, \ \forall \bar{y} \neq y^{(i)}: \tag{6}$$

$$\langle w, \Psi(x^{(i)}, y^{(i)}) - \Psi(x^{(i)}, \bar{y}) \rangle \geq \Delta(y^{(i)}, \bar{y}) - \xi^{(i)},$$

$$\forall i: \quad \xi^{(i)} \geq 0.$$

The above minimization problem has one constraint for each image. Every constraint is associated with a slack-variable $\xi^{(i)}$ that acts as an upper bound on the error Δ caused by annotating the ith image with a wrong label. Once, optimal parameters w^* have been found, these are used as plug-in estimates to compute predictions for new and unseen examples

using (4). The computation of the argmax can be performed by explicit enumeration of all paths in the taxonomy.

An alternative formulation (Tsochantaridis et al. 2005) uses slack rescaling instead of margin rescaling in the constraints:

$$\min_{w,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi^{(i)}$$

$$\text{s.t.} \quad \forall i, \ \forall \bar{y} \neq y^{(i)}:$$

$$\langle w, \Psi(x^{(i)}, y^{(i)}) - \Psi(x^{(i)}, \bar{y})\rangle \geq 1 - \frac{\xi^{(i)}}{\Delta(y^{(i)}, \bar{y})},$$

$$\forall i: \quad \xi^{(i)} \geq 0. \tag{7}$$

In this multiplicative formulation based on a hinge loss

$$\max\Big(0, \max_{\bar{y}} \Delta(\bar{y}, y^{(i)})$$

$$\times \langle w, \Psi(x^{(i)}, \bar{y}) - \Psi(x^{(i)}, y^{(i)})\rangle\Big) \tag{8}$$

each sample receives the same margin of one. As a drawback finding the maximally violated label can be more complicated compared to margin rescaling due to the label $\bar{y}$ appearing in both factors of a product. Margin rescaling is also based on the hinge loss but uses an additive formulation in $\Delta(\bar{y}, y^{(i)})$

$$\max\Big(0, \max_{\bar{y}} \Delta(\bar{y}, y^{(i)})$$

$$+ \langle w, \Psi(x^{(i)}, \bar{y}) - \Psi(x^{(i)}, y^{(i)})\rangle\Big) \tag{9}$$

where it might be easier to find the maximally violated constraint but on the other side here the loss function Δ might dominate the loss term (9) if it is badly scaled.

Although, (6) and (7) can be optimized with standard techniques, the number of categories in state-of-the-art object recognition tasks can easily exceed several hundreds which renders the structural approaches inherently slow.

2.3 Assembling Local Binary SVMs

We propose here an efficient alternative to the structural approaches by decomposing the structural approach from (6) into several local tasks. The idea is to learn a binary SVM (e.g. Cortes and Vapnik 1995; Müller et al. 2001; Schölkopf and Smola 2001) using the original representation $\phi(x)$ for *each node* $v_j \in V$ in the taxonomy instead of solving the *whole* problem at once with a structured learning approach. This will help to circumvent the high computational load typically encountered in structured learning. To preserve the predictive power, the final ensemble of binary SVMs from each node need to be assembled in an intelligent manner, i.e. appropriately according to the taxonomy.

We remark that this novel approach is different from greedy hierarchical classifiers where at each node only categories (leaf nodes) below it are taken into account. On the contrary, we are considering *all* images and categories at each node: for example, we learn binary SVMs such as 'Carnivora vs the others' and 'horse vs the others', while only 'Carnivora vs horse', 'cat vs dog' etc. would be used in the greedy hierarchical classification. As outlined in Sect. 4.7.2, the greedy approaches perform sub-optimally, because they may rely on erroneous decisions of upper internal nodes.

Thus essentially, our approach consists of training $|V|$ independent binary support vector machines (which can be done highly efficiently in parallel!) such that the score $f_j(x) = \langle \tilde{w}_j, \phi(x)\rangle + \tilde{b}_j$ of the j-th SVM centered at node v_j serves as an estimate for the probability that v_j lies on the path y of instance x, i.e., $Pr(\kappa_j(y) = 1)$. An image $x^{(i)}$ is therefore treated as a positive example for node v_j if this very node lies on the path from the root to label $y^{(i)}$ and as a negative instance otherwise, which amounts to the sign of $2\kappa_j(y^{(i)}) - 1$.

We resolve our *local-SVM* optimization problem that can be split into $|V|$ independent optimization problems, effectively implementing a one-vs-all classifier for each node.

$$\min_{\tilde{w}_j, \tilde{b}_j, \tilde{\xi}_j} \quad \frac{1}{2}\sum_{j=1}^{|V|}\|\tilde{w}_j\|^2 + \sum_{j=1}^{|V|}\tilde{C}_j\sum_{i=1}^{n}\tilde{\xi}_j^{(i)}$$

$$\text{s.t.} \quad \forall i, \ \forall j: \quad (2\kappa_j(y^{(i)}) - 1)(\langle \tilde{w}_j, \phi(x^{(i)})\rangle + \tilde{b}_j)$$

$$\geq 1 - \tilde{\xi}_j^{(i)},$$

$$\forall i, \ \forall j: \quad \tilde{\xi}_j^{(i)} \geq 0. \tag{10}$$

At test phases, the prediction for new and unseen examples can be computed similarly to (4). Denote the local-SVM for the j-th node by f_j, then the score for class y is simply the sum of all nodes lying on the path from the root to the leaf y,

$$f_y(x) = \frac{\sum_{j:\kappa_j(y)=1} f_j(x)}{\sum_j \kappa_j(y)}. \tag{11}$$

The normalization is required due to varying path lengths in our taxonomies which is a difference compared to the taxonomies considered in Cai and Hofmann (2004). The class y which has the maximum score f_y over all classes is selected as the final prediction.

Note that since the entire problem decomposes into $|V|$ binary classification tasks, parallelization becomes possible and thus, the training time of our approach is considerably shorter compared to the structural SVMs. Another advantage is that our local procedures can be directly extended to multi-label problems without taking the maximum operation at the end, but by setting thresholds only which determine whether object categories are included in images or not.

Although our initial motivation was to construct an efficient approximation of the structural SVMs, we would like to remark that there exists a fundamental difference between the structural SVMs and our local-SVM procedure with respect to their optimization target. The constraints of the structure learning in (6) aim to order the *set of all class labels* correctly *for each image* in the sense that the SVM score for the correct class label is highest. For our local-SVM approach the SVM constraints aim at ordering the *set of all images* correctly *for each node* with respect to the binarized learning problem whether an image belongs to a class lying on a path passing through this taxonomy node or not. We remark further that the constraints of the structural optimization problems do not imply necessarily that the set of all images is ordered correctly for the binary classification problem at each taxonomy node. In order to foster a better intuitive understanding, the difference between both approaches are illustrated in Fig. 3.

2.4 Scoring with Generalized p-means

When we combine the binary classification scores at the nodes along a path, it is not necessary to take their arithmetic mean as in (11). Instead, our procedures permit more general scoring methods such as the generalized p-means of outputs

$$M_p(z_1, \ldots, z_m) = \left(\frac{1}{m} \sum_{i=1}^{m} z_i^p \right)^{1/p}, \tag{12}$$

after scaling to $[0, 1]$. This includes the geometric mean as the limit $p \to 0$ and the harmonic mean for $p = -1$ as well as the minimum as the limit $p \to -\infty$. Tuning of this extra degree of freedom p may improve classification performance. To see this note that the geometric mean and generalized means with negative norms of scores in $[0, 1]$ are upper bounded by a power of the smallest element.

$$s_i \in [0, 1] \Rightarrow \prod_{i=1}^{n} s_i^{1/n} \leq \min_i s_i^{1/n},$$

$$p < 0 \Rightarrow \left(\frac{1}{n} \sum_{i=1}^{n} s_i^p \right)^{1/p} \leq \frac{1}{n^{1/p}} \min_i s_i.$$

For positive norms the generalized mean is upper bounded instead by a power of its largest element. In that sense generalized means with non-positive norms are more sensitive to negative outliers and more robust against strong positive outlier votes from nodes than generalized means with positive norms where the distortion by strong positive outliers can be arbitrarily large. The selection of an optimal p-norm thus adjusts the sensitivities to very small votes close to 0 versus very large votes close to 1. The usage of generalized
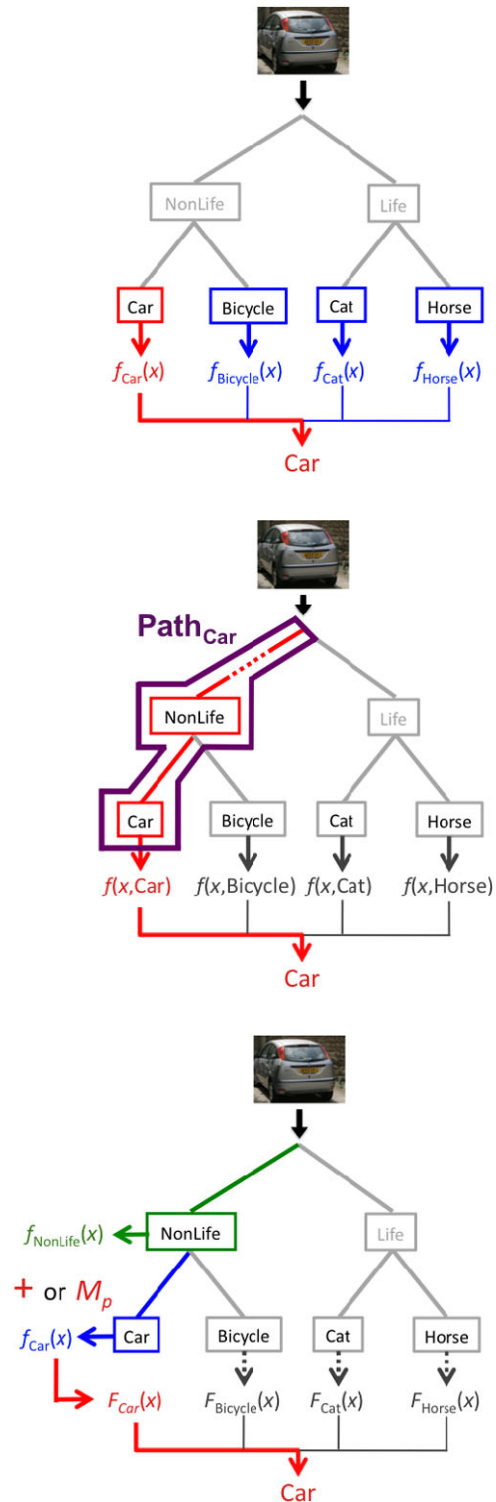

Fig. 3 Differences between one vs all (*top*), structure learning (*middle*) and local approach (*bottom*). The one vs all procedure ignores internal nodes of taxonomies and takes the maximum of the SVM outputs at leaf nodes. The structured approach takes paths as a whole into account, maximizes the margin between correct and wrong paths in training and returns as a predictor the label of the path with the maximum score. The local procedures optimize each binary problem of passing through a path independently and then combine the outputs of the local SVMs into a score with generalized p-means

means with arbitrary norms requires the scores to be non-negative and SVM outputs to be scaled.[4]

In order to scale SVM outputs into $[0, 1]$, we deploy a logistic function with fixed parameters

$$s(y) = \frac{1}{1 + \exp(-10y)}.$$

Experimentally we have seen that learning the logistic regression parameters from data (Platt 1999) did not further improve performance of image categorization.

Scaling with logistic functions is closely linked to a probabilistic interpretation of a classification procedure. While this is common for greedy hierarchical classification, our current approach does not immediately permit a probabilistic interpretation fitting to a taxonomy graph. This is because we so far have chosen to always consider classification between a part of the categories and all remaining others at each node, instead of conditioning on its parent nodes, for efficiency reasons.

2.5 Baselines

In our experiments, we will use additionally two kinds of classification methods. One is the standard one-vs-all classification: we train one binary SVM for each class which uses the samples of this class as positive labeled data and all the other class data as negative examples. The multi-class labeling is obtained by the class maximizing the scores of all binary SVMs. This is a completely taxonomy-free approach. The second is structured multi-class SVMs which uses the joint feature representation ignoring the taxonomy graph

$$\Psi(x, y) = \phi(x) \otimes \iota(y) = \begin{pmatrix} \phi(x)[[y = c_1]] \\ \phi(x)[[y = c_2]] \\ \vdots \\ \phi(x)[[y = c_k]] \end{pmatrix},$$

where $\iota(y)$ is the vector of the indicator functions $[[y = c_i]]$. This leads to the 0/1 loss from the label kernel

$$2 - 2K_Y(y_1, y_2) = \delta_{0/1}(y_1, y_2),$$

instead of the taxonomical one in the structured taxonomical SVMs. No taxonomy information is used, if the 0/1 loss is deployed as the loss function Δ in (6) and (7), while it is incorporated indirectly into the learning process, when Δ is the taxonomy loss δ_T.

---

[4]While there exist convex mappings of $\mathbb{R}^1$ to the interval $[0, \infty)$ we are not aware of the existence of a monotonous and continuous mapping of $\mathbb{R}^1$ onto a bounded nontrivial interval which is everywhere concave or convex. This implies that a model using scaling of unbounded inner products cannot be optimized by applying convex methods in the structured output framework.

3 Insights from Synthetic Data

In this section, we discuss when and why the taxonomical approaches might outperform the one-vs-all baseline. Furthermore we can observe differences in AUC scores between leaf and internal nodes which can be linked to flat losses in later experiments on real data. We remark that the one-vs-all baseline can be regarded as a classification procedure only with leaf nodes, while the taxonomy-based learning combines classification results of leaf and internal nodes, namely by generalized p-means in the local-SVM approach and by implicit arithmetic mean integrated in the structural SVMs.

3.1 Experimental Results

To illustrate our claim, we consider a 16 class example with the taxonomy being a binary balanced tree with 16 leaf nodes. Each class is generated from one Gaussian distribution in 15 dimensions. The variances are equal for all Gaussian and are varied to give seven datasets with $\sigma = 1$, 0.5, 0.3725, 0.25, 0.1875, 0.125, 0.0625. The means are distributed such that their Euclidean distance matrix equals the normalized taxonomy loss matrix which has values $i/4$, $i = 0, \ldots, 4$. Our intention is to illustrate that taxonomy-based learning reduces taxonomy loss, if the data is aligned to the taxonomy. For the sake of computation speed we compare the one-vs-all baseline versus a local algorithm with scoring based on the geometric mean of logistically scaled scores of 19200 data points each independently, where we use 200 samples per class for training and the remaining 1000 per class for testing. We deployed Gaussian kernels here, set the width to be the mean of squared distances and normalized all kernels to have standard deviation one in Hilbert space.

Table 1 shows the 0/1 and taxonomy losses of one-vs-all and our local SVM procedure with the scaled geometric mean over different noise levels. The standard deviations are computed between the 15 draws.

The local algorithm improved the one-vs-all baseline significantly under the taxonomy loss in all cases. The relative

Table 1 Synthetic data perfectly aligned to the taxonomy: Losses of the one-vs-all baseline (left) versus local procedure with taxonomy (right) for different label noise levels

σ	One-vs-all		Local-SVM approach	
	$\delta_{0/1}$	δ_T	$\delta_{0/1}$	δ_T
1	89.10 ± 0.32	67.09 ± 0.34	$\mathbf{88.59} \pm 0.34$	$\mathbf{65.69} \pm 0.35$
1/2	78.24 ± 0.32	51.37 ± 0.31	$\mathbf{77.84} \pm 0.39$	$\mathbf{50.27} \pm 0.35$
3/8	69.30 ± 0.38	41.29 ± 0.28	$\mathbf{68.94} \pm 0.39$	$\mathbf{40.21} \pm 0.29$
1/4	51.61 ± 0.52	25.05 ± 0.26	$\mathbf{51.26} \pm 0.52$	$\mathbf{24.17} \pm 0.22$
3/16	37.32 ± 0.46	14.94 ± 0.23	$\mathbf{36.91} \pm 0.48$	$\mathbf{14.24} \pm 0.23$
1/8	19.49 ± 0.39	6.05 ± 0.11	$\mathbf{19.12} \pm 0.41$	$\mathbf{5.70} \pm 0.12$
1/16	2.41 ± 0.13	0.61 ± 0.03	$\mathbf{2.38} \pm 0.13$	$\mathbf{0.60} \pm 0.03$

Table 2 Synthetic data perfectly aligned to the taxonomy: AUC scores in the taxonomy for $\sigma = 1/4$ at different levels

Level in taxonomy	1	2	3	4 (leaf)
AUC	99.21	97.78	95.42	92.40

Table 3 Synthetic data perfectly aligned to the taxonomy: At which level does misclassification occur for $\sigma = 1/4$?

Level in taxonomy	1	2	3	4 (leaf)
Differences of Error Rates	−1.55	−0.68	0.48	1.74

improvements are more than 2% with the maximum above 5% for $\sigma = 1/8$. We also conducted Wilcoxon's signed rank test, which showed that all performance gains are significant with p-values of orders 10^{-4} or 10^{-5}. Surprisingly, the local SVM procedure the taxonomy compares favorably with the baseline under the flat 0/1 loss as well.

There is an intuitive explanation why hierarchical approaches do improve losses consistent with the hierarchy compared to one versus all classifiers. One versus all classifiers attempt to rank the images belonging to positive class highest. Classifiers from superclasses in a hierarchy attempt to rank the images belonging to the positive class *and similar classes* to be highest. Averaging many versus all classifiers from superclasses with one versus all classifiers at the leafs achieves a tradeoff between both aims. At the same time such an averaging can potentially harm the zero-one-loss which does not consider similarities encoded in a taxonomy.

Table 2 shows the AUC score at different levels in the hierarchy. It allows to judge how difficult the learning problems are at the internal nodes compared to leaf nodes. Note that we observe on this synthetic dataset a higher AUC score on internal nodes compared to leaf nodes and a decrease in the flat zero-one-error compared to the one versus all baseline. This implies that the learning problems are easier on superclass level than at the leaf nodes. This might explain why we observe here an improvement in the flat zero one loss as well. It is not straightforward in a statistical sense that optimizing for one loss improves another loss as well. As an explanation we propose that in this synthetic case the features allow a good generalization at superclass level because the given taxonomies are perfectly aligned to the similarities between classes at the feature level. The higher AUC score at internal nodes compared to leaf nodes supports this view. This good alignment might be also the case when learning similarities from visual features and explain results for flat losses in Marszalek and Schmid (2008), Zweig and Weinshall (2007) but it cannot be expected to hold in general when a taxonomy is provided independent of visual features. We will return to this observation in the forthcoming Sect. 4 on experiments on real data.

Table 3 shows another aspect of hierarchical averaging: given a pair consisting of true and predicted label we can ask where in the hierarchy the error did occur. This leads to two histograms, for the taxonomy-based and for the one versus all classifier. The table shows the difference between both histograms. Negative values imply a reduction of errors at this level for the taxonomic method. We see that under our taxonomy based approach the classification errors are moved to lower levels in the hierarchy compared to a flat one versus all classification implying that confusions occur more often between taxonomically closer classes.

3.2 Robustness by p-means

The parameter p of the generalized controls robustness against outlying classifier outputs. Negative p's make the mean robust against upper extremes while in the opposite cases lower extremes are suppressed. To see this we conducted an experiment on controlled perturbation of SVM outputs over the toy data. We fixed a priori a set of 10% of the samples to be perturbed and for each sample one node in the taxonomy to be perturbed. We applied these fixed sets to values of perturbation factors $\{+8, +4, -4, -8\}$. The perturbation is computed for a sample by adding to the SVM output of this sample the factor times the standard deviation of the outputs of the SVM corresponding to the taxonomy node. The negative factors allow to simulate large negative outliers, the positive factors large positive outliers. Table 4 shows the results.

We can see that for large positive distortions both positive means perform lower than geometric mean and a negative mean.

For large negative distortions the first ranks are held by the non-scaled arithmetic mean and a scaled positive mean. These two methods suffer less from negative outliers than negative means. Furthermore we observe in both settings that unscaled variants are less robust than scaled ones.

Finally the last part of the Table 4 shows a result where 80% of the perturbed samples are modified by a factor of $+4$ and 20% by -4. Here the geometric mean turns out to be the best choice which corresponds well to our empirical findings in Sect. 4.5. We conclude that the geometric mean is well suited to deal with SVM outputs which suffers from positive and negative outliers in taxonomy nodes coming from noisy classification problems.

In summary, we would like to emphasize that classification techniques with taxonomies can improve the one-vs-all baselines, under the taxonomical loss and the flat zero one loss.

Table 4 Synthetic data perfectly aligned to the taxonomy: Differences in taxonomy loss and 0/1 loss to unperturbed SVM outputs and absolute ranks between all four methods

Unperturbed	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Rank δ_T	**1**	3	2	4
Rank $\delta_{0/1}$	**1**	3	**1**	4
Perturb = +8	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Diff. δ_T	1.8	0.14	0.04	0.05
Rank δ_T	4	3	**1**	2
Diff. $\delta_{0/1}$	1.91	0.27	0.15	0.15
Rank $\delta_{0/1}$	4	3	**1**	2
Perturb = +4	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Diff. δ_T	0.47	0.14	0.04	0.05
Rank δ_T	4	3	**1**	2
Diff. $\delta_{0/1}$	0.81	0.26	0.15	0.15
Rank $\delta_{0/1}$	4	3	**1**	2
Perturb = −4	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Diff. δ_T	0.26	0.03	0.42	0.75
Rank δ_T	2	**1**	3	4
Diff. $\delta_{0/1}$	0.34	0.13	0.49	0.73
Rank $\delta_{0/1}$	**1**	2	3	4
Perturb = −8	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Diff. δ_T	0.68	0.03	0.7	0.75
Rank δ_T	2	**1**	3	4
Diff. $\delta_{0/1}$	0.73	0.12	0.74	0.74
Rank $\delta_{0/1}$	2	**1**	3	4
80% +4, 20% −4	Nonscaled M_1	sc M_2	sc M_0	sc M_{-2}
Diff. δ_T	0.41	0.09	0.11	0.12
Rank δ_T	4	3	**1**	2
Diff. $\delta_{0/1}$	0.53	0.21	0.2	0.23
Rank $\delta_{0/1}$	4	3	**1**	2

4 Experiments on Real World Multiclass Data

4.1 Datasets

For the present work, we constructed multi-class classification datasets with taxonomy trees between object categories by modifying the benchmarks Caltech256 (Griffin et al. 2007) and VOC2006 (Everingham et al. 2006).

Caltech256 all classes The Caltech256 dataset (Griffin et al. 2007) contains 256 classes of objects and one clutter class. For an initial experiment allowing comparison to results from other publications we have taken 50 images from each of the object classes and employed the taxonomy as provided in the report (Griffin et al. 2007). The only changes we made were to add pisa-tower to the taxonomy graph as it seemed to be missing and moved iris to flowers from air

animals. Unfortunately, using $50 \cdot 256 \cdot 0.9 = 11520$ samples for training using ten-fold crossvalidation is beyond the scope of the structured prediction baselines on our hardware. Therefore we considered subsets of classes which will be described below. The result for all 256 object classes can be looked up in Sect. 4.7.5.

Caltech256 animals We consider all 52 real world animal classes from the Caltech256 dataset (Griffin et al. 2007) which yields 5895 data points (see Fig. 4). They form a multi-class problem with mutually exclusive classes. We used a taxonomy based on a recherche of biological (phylogenetic) systematics consisting out of 92 nodes constructed a priori. We have chosen this subset for two reasons. Firstly, it is a natural multiclass dataset in the multimedia image domain. Secondly, it allows to define a taxonomy in an undisputable way prior to looking at image content, namely via biological systematics. For the remaining 204 classes from Caltech256 we would have to rely on human experience of some sort which might lead to some kind of unintentional appearance-based optimization of when choosing a taxonomy. The technical report on the Caltech256 dataset (Griffin et al. 2007) contains a hierarchy. We have chosen not to use its construction principle because it is somewhat arbitrary as stated by the authors of the technical report themselves and from our own point of view is not biologically plausible. It groups all animals in four flat subgroups: insects, land, air and water based lifeforms. As stated in the introduction the usage of phylogenetic systematics resulted in a taxonomy which is indeed not fully consistent to the subjective visual similarities of the authors which diverge for example for crabs and horseshoe crabs but also as shown in Fig. 1 potentially for superclasses in the taxonomy. The hierarchy contains in contrast to many preceding works paths with varying lengths. We omitted fantasy animals like Minotaurs and Unicorns from the Caltech256 set, as there is no objective way to incorporate them into biological systematics. The full taxonomy is given in Fig. 10 in the Appendix.

Caltech256 animals thirteen classes subset For further experiments, we select 13 classes—all Protostomia (prayingmantis, grasshopper, cockroach, house-fly, butterfly, trilobite, centipede, crab, spider, scorpion, horseshoe-crab, octopus, snail) from the *Caltech256 animals* dataset. This corresponds to one subtree in the original taxonomy over all 52 classes. The total number of the images is reduced to 1308. This allows us faster experimentation with the structural approaches which was the main reason for choosing this subset. We deploy as taxonomy the corresponding subtree with 21 nodes of that of *Caltech256 animals* which is still challenging in its topology due to non-balanced tree structure and varying path lengths.

VOC2006 multi-class data We use the VOC2006 dataset (Everingham et al. 2006) consisting of 10 object classes and 5301 images (see Fig. 5). We have modified the VOC2006 labels in order to obtain a multi-class problem with mutually exclusive classes. To achieve such exclusive labeling, for each image all positive labels except for a randomly chosen one are suppressed. We remark that this process induces additional label noise.

4.2 Image Features

For the following experiments, we used bag of words (BoW) representations based on the SIFT descriptors (Lowe 2004) as image features. The BoW features were constructed in a standard way: using the code from van de Sande et al. (2010), the SIFT descriptors were computed on a dense grid of step size six over the color channel triples {red, green, blue} and {grey, opponent color 1, 2}. Then, for both triples, 8192 visual words (prototypes) were generated by using ERCF clustering (Moosmann et al. 2008) via 16 trees with 512 leaves each based on large sets of SIFT descriptors selected randomly from the training images following (van de Sande et al. 2010). For each image, each SIFT feature was assigned to one leaf for each of the 16 trees. We have chosen the supervised ERCF procedure over k-means as it does greatly reduce the time necessary for clustering of visual words and bag of word computation while having comparable performance. The sum of these mappings resulted in one histogram for each image within each cell of the spatial tilings 1×1, 2×2 and 3×1 (Lazebnik et al. 2006; Bosch 2007). Finally, we obtained 6 BoW features (2 colors $\times$ 3 pyramid levels) with dimensionalities 8192, 4×8192 and 3×8192 depending on the spatial tiling. For Caltech 256 data we omitted the two kernels based on tilings 2×2 as they did degrade the one-vs-all baseline performance already. We do not aim here at the best possible baseline performance which might be achieved by adding carefully selected sets of additional features. Instead we focus on the effect of a given hierarchy and non-flat loss functions. We note however that high-dimensional bag of words models have been able to achieve superior performance in recent object categorization challenges (van de Sande et al. 2010; Everingham et al. 2007, 2008, 2009; Tahir et al. 2008) which motivates our choice of these features.

4.3 Image Kernels and Regularization of SVMs

We used the Chi2-Kernel for comparing the image feature histograms (Zhang et al. 2007)

$$k_\sigma(v, w) = \exp\left(-\sigma \sum_{d|w_d+v_d>0} \frac{(w_d - v_d)^2}{2(w_d + v_d)}\right).$$

The kernel width was fixed to be the mean of all Chi2-distances. All kernels have been normalized to standard deviation in Hilbert space set equal to one which in practice limits the range where to search for an optimal regularization constant. We combined all kernels via addition.

In the local-SVM procedure, we used two regularization constants (one per class) for all binary problems in order to compensate for the unbalanced ratios between positive and negative classes. The regularization constant of the smaller class was obtained by multiplying that of the larger class[5] by the ratio between the two samples. For the structured SVMs we used as regularization parameter $\tilde{C} = 16|V|$ for the taxonomical procedures and $\tilde{C} = 16k$ for the multi-class ones, where $|V|$ and k are the number of nodes and classes, respectively.

This is motivated by comparing the main objective of one local SVM

$$\min_{\tilde{w}_j, \tilde{b}_j, \tilde{\xi}_j} \quad \frac{1}{2}\|\tilde{w}_j\|^2 + \tilde{C}_j \sum_{i=1}^n \tilde{\xi}_j^{(i)}$$

to the one from a structured SVM

$$\min_{\tilde{w}, \tilde{\xi}} \quad \sum_{j=1}^{|V|} \frac{1}{2}\|\tilde{w}_j\|^2 + \tilde{C} \sum_{i=1}^n \tilde{\xi}^{(i)}.$$

We note that the ratio between the weight norm $\|w\|^2$ and the slacks $\xi^{(i)}$ is roughly up-scaled by a factor equal to the number of nodes. We have checked experimentally that using much lower regularization constants damages the performance of the structural SVMs, while much higher regularization constants did not improve the results anymore. Since the sizes of the object categories are balanced, we do not have to assign one regularization constant for each class separately.

4.4 Comparison Methodology

All considered methods can be divided into structured and structure-free as well as taxonomical and taxonomy-free approaches (Table 5). Due to limited space, we will use the abbreviations listed in Table 6 to in our experimental results.

There are three ways to use the taxonomy. The taxonomy loss as performance measure is used on all methods. The taxonomy loss as part of the training procedure is used in all structured SVMs according to (6). The taxonomy structure is incorporated in all taxonomical approaches but not in the structured multi class procedures.

---

[5]The regularization constant of the larger class was fixed to 16 which corresponds to our experience that high-dimensional Bag-of-words features perform better under hard margin training.
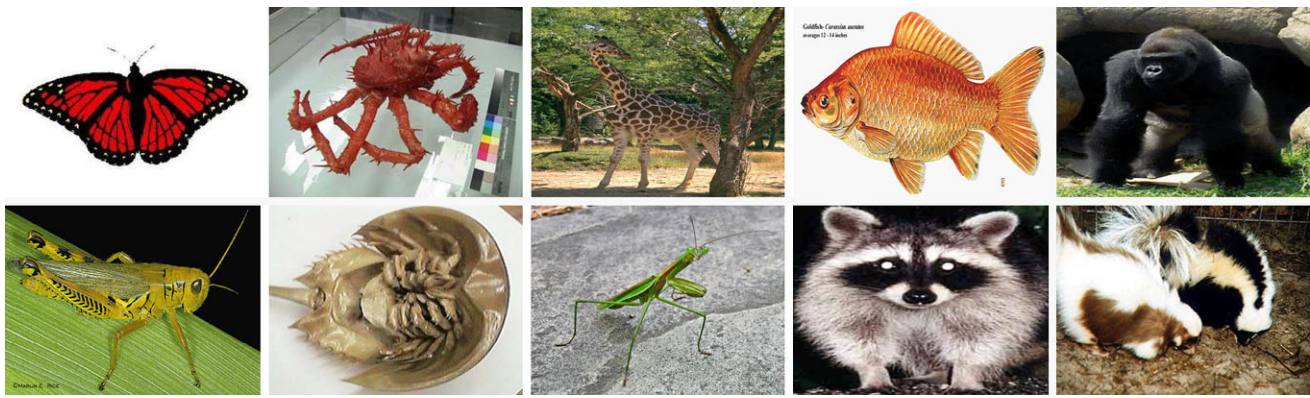
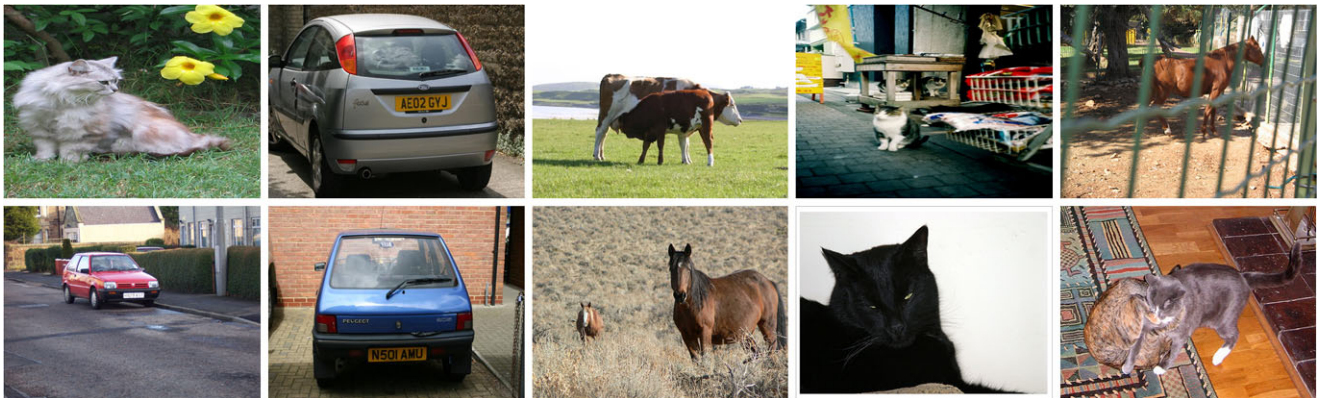**Fig. 4** Caltech 256 animals example images

Fig. 5 VOC2006 example images

Table 5 Classification of methods

	Structure-free	Structured
Taxonomy-free	One vs all	Struct multi-class SVMs
	(Sect. 2.5)	(Sect. 2.5)
Taxonomical	Local taxonomy	Struct taxonomy SVMs
	(Sect. 2.3)	(Sect. 2.2)

Table 6 Abbreviations for compared methods

Structured multi-class baseline	
Struct mc mr	With margin rescaling
Struct mc sr	With slack rescaling
Taxonomical structural learning	
Struct tax mr	With margin rescaling (6)
Struct tax sr	With slack rescaling (7)
The local procedure with taxonomy	
Local tax AM	With arithmetic mean (11)
Local tax scaled GM	With geometric mean after scaling
M_p	With p-mean after scaling

We will use as baselines the structure-free one-vs-all classification and taxonomy-free multi class SVMs with margin and slack rescaling trained using zero-one loss $\delta_{0/1}$ or taxonomy loss δ_T. The taxonomy-based algorithms to be tested consist of the structured SVMs with nontrivial taxonomies in margin (6) and slack rescaling formulation (7) and of structure-free methods scoring via the arithmetic mean over the component SVM outputs and via generalized means of them scaled using logistic functions.

We used SVMmulticlass (Joachims 1999) and modified versions thereof for the structured approaches. The non-structured methods have been implemented using shogun toolbox (Sonnenburg et al. 2010) with the SVMlight solver.

We note that SVMlight is also deployed in the optimization procedures of the SVMmulti-class implementations.

The error measurement is done for the multi-class problems using the 0/1- and taxonomy loss from (2). For all multi-class problems we use 20 splits into training and test data with 50 images per class in each split. This greatly reduces the dataset size compared to cross-validation and the training time for the structured methods.

Table 7 One-vs-all performance on multi-class datasets

Dataset	0/1 loss	AP score
Cal256 animals	62.56	34.34
Cal256 13 class subset	57.04	43.69
VOC2006, multi-class, 20 splits	50.54	54.75
VOC2006, multi-class, 20-fold crossval	33.56	70.50

4.5 Experimental Results: Performance Comparisons

At first, we would like to remark the difficulty inherent in the datasets. Table 7 shows the 0/1 loss and the average precisions (AP score) of the one-vs-all baselines for the three multi-class datasets.

The AP score is a rank-based measure which was deployed as the performance criterion in the recent Pascal VOC challenges. For VOC2006 the results for 20 splits perform worse due to sample size effects as they use only 500 training data in each split as compared to over 5000 points for the 20-fold cross-validation.

The comparisons for Caltech256 animals and its 13 class subset are shown in Tables 8 and 9. For simplicity, we present only the best result among all options for each of structural multi-class, local taxonomy-based and structural taxonomy-based procedures. The full Tables listing all results can be found in Appendix (Tables 17–19). As expected, the taxonomy-based methods outperform the taxonomy-free baselines in terms of the taxonomy loss by 3–5% relatively. For both datasets, our local SVM procedure improves structure learning with taxonomy by 2–3% relatively. The gains of the taxonomy-based approaches under the taxonomy loss are achieved at the cost of slightly increasing the 0/1 loss. It is notable from Table 9 that merely including the taxonomy loss in a structured multi-class algorithm (as an intermediate step of incorporating taxonomical information) does not yield sufficient performance gain under the taxonomy loss. Optimization for taxonomy loss comes at the cost of performance deterioration under the 0/1 loss. This is not surprising, because the baselines, one vs all and structured multi-class models directly optimize for the flat hinge loss which is more closely related to the 0/1 loss than to the taxonomy loss. Since this problem occurs for all hierarchical methods including the structured prediction based methods it does point out the considerable difference between the canonical flat loss and what a user might desire. From an optimization viewpoint minimizing a different loss leads to a different model. Therefore merely the scale of change might be surprising. The relation of 0/1 loss to AUC scores at internal nodes across datasets will be discussed in Sect. 4.7.3.

Table 10 shows the performance comparison for the VOC2006 multi-class problem. Similar to the Caltech animals datasets, the taxonomy-based methods outperform the one-vs-all baseline in terms of the taxonomy loss by 5%

Table 8 Performance on Caltech256 animals (52 classes), 20 splits

Method	Taxonomy loss	0/1 loss
One vs all	30.66 ± 0.46	$\mathbf{62.56} \pm 0.67$
Best local tax: scaled GM	$\mathbf{29.62} \pm 0.34$	76.19 ± 0.57
Best struct tax: mr	30.58 ± 0.31	81.19 ± 0.53

Table 9 Performance on Caltech256 animals 13 class subset data, 20 splits

Method	Taxonomy loss	0/1 loss
One vs all	42.49 ± 1.46	$\mathbf{57.04} \pm 1.98$
Best struct mc: sr, $\Delta = \delta_{0/1}$	42.48 ± 1.50	57.06 ± 2.00
Best local tax: scaled GM	$\mathbf{40.58} \pm 1.15$	58.33 ± 1.50
Best struct tax: mr	41.48 ± 1.22	61.54 ± 1.55

Table 10 Performance on VOC2006 as multi-class problem, 20 splits

Method	Taxonomy loss	0/1 loss
One vs all	27.09 ± 1.88	50.54 ± 2.51
Best struct mc: mr, $\Delta = \delta_T$	26.37 ± 1.77	51.04 ± 2.53
Best local tax: scaled GM	25.86 ± 1.56	$\mathbf{50.10} \pm 2.29$
Best struct tax: mr	$\mathbf{25.78} \pm 1.67$	50.17 ± 2.17

relatively. On the other hand, there are some differences from the previous cases. At first, our local SVM procedure were rather on par with the structural counterpart. Secondly, the intermediate step, the structure multi-class procedure with the taxonomy loss δ_T improved the one-vs-all baseline significantly under the taxonomy loss. Finally, the taxonomy-based approaches improved slightly the taxonomy-free baselines under the 0/1 loss as well as was the case for the synthetic example.

As a sanity check for structured implementations we remark that the structure-free methods perform approximately equally well to their structured counterparts for both taxonomy and 0/1 losses. Since for the flat 0/1 loss setting we used SVMstruct in its unmodified formulation, this is clearly a property of the data rather than a potentially faulty implementation of structured approaches.

In summary, we observed that the taxonomical approaches outperform the taxonomy-free baselines under the taxonomy loss, as was the case for the synthetic data. Unlike in the synthetic data the zero-one error was slightly increased by optimization of taxonomy based losses for both Caltech datasets. The choice of the loss function determines the algorithm to be used. It is not expectable in a statistical sense that a taxonomical model improves a flat loss under all circumstances, however there is a tendency for relatedness of zero one loss and differences of AUC scores

(see also discussion in Sect. 4.7.3). The local taxonomy-based methods are slightly worse than structured taxonomy ones on VOC2006 dataset, but considerably better on both Caltech256 animals problems. We would like to emphasize that the way of averaging is important to achieve better performance. Note that the scaled geometrical mean compares favorably with the arithmetic mean. Indeed, when we examined the generalized p-means in a wide rage of the parameter p, parameters close to 0 (i.e. the geometrical mean) achieved the minimum values both under the 0/1 and taxonomy losses.

4.6 Training Time

In all three data sets the local SVMs are much faster to train when compared to structured taxonomy approaches (cf. Table 11). The local SVMs can be parallelized by training each node as a separate optimization problem, an advantageous property when scaling the number of object categories. Another beneficial scaling characteristic when increasing the number of samples is the possibility to reduce the training set for each node individually since it is sufficient to control the performance of the binary classification problem at each node separately. Certain steps in the structural approaches like finding the most violated constraints can be parallelized to e.g. multicore machines which typically accounts for four or at most eight cores. The used code may have potential for further problem-specific optimizations. The speed gains by using local SVMs are large factors of over 10. Thus we do not expect the advantage of the local svms to disappear against a multicore-parallelization of structural support vector machines. Furthermore the parallelization of local svms into optimization problems restricted to single nodes can be achieved generically over more than 8 cores. Another performance reducing factor was excessive main memory usage of

structural algorithms of up to 16 Gigabyte per task which in practice leads to additional slowdowns compared to many small tasks as solved by the local SVMs.

4.7 Discussion

4.7.1 Confusion Between Object Categories

Figures 7 and 8 provide example images where the results from the local taxonomy approach differs compared to the one versus all baseline. Each image comes with a graph on the taxonomy. The ground truth label is green. The choice by one versus all is marked in magenta and the path to the choice by hierarchical classification is given in blue. All relevant paths have attached the SVM outputs to them (see also Fig. 3). Figure 7 shows typical cases when the hierarchic approach fails. It is caused by false positive outlier

Table 11 Training times, the multiplier for local models shows separability into independent jobs

| Method	Dataset	Training time
One vs all	Cal256 animals, 52 classes	3.69 s × 52
Local tax	Cal256 animals, 52 classes	3.69 s × 92
Struct. tax	Cal256 animals, 52 classes	35.13 h
One vs all	Cal256 animals, 13 classes	0.5 s × 13
Local tax	Cal256 animals, 13 classes	0.5 s × 21
Struct multi-class	Cal256 animals, 13 classes	15.1 min
Struct tax	Cal256 animals, 13 classes	44.9 min
One vs all	VOC2006	<0.5 s × 10
Local tax	VOC2006	<0.5 s × 19
Struct multi-class	VOC2006	9.4 min
Struct tax	VOC2006	28.7 min
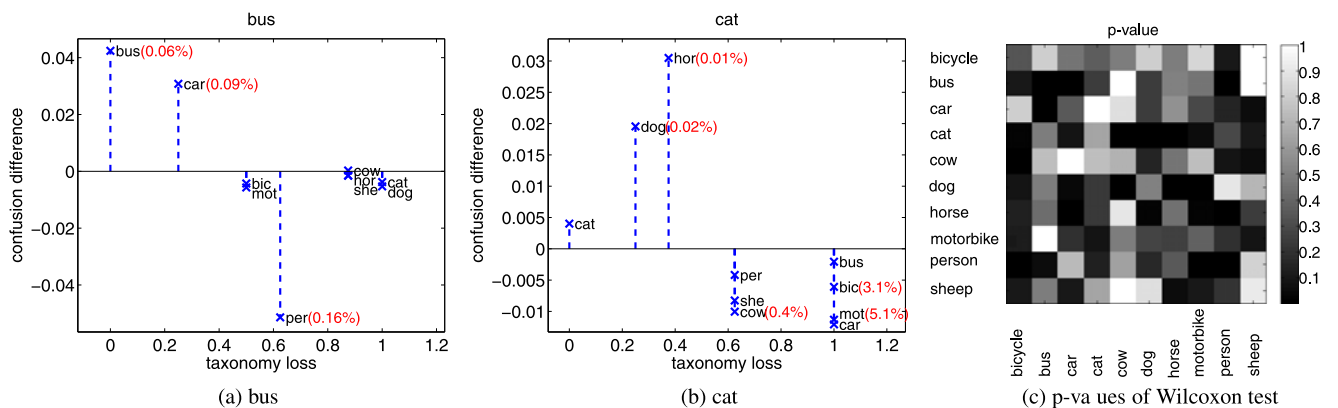
Fig. 6 Confusion differences between our local SVM with taxonomy and the one-vs-all classification (y-axis) versus the taxonomy losses (x-axis) for (**a**) bus and (**b**) cat from VOC 2006 categories (bic = bicycle, hor = horse, mot = motorbike, per = person, she = sheep). Positive values denote more confusions by the proposed method. Significances of the differences are checked by Wilcoxon signed-rank test whose p-values are summarized in (**c**) (row: true classes, column: predicted classes)
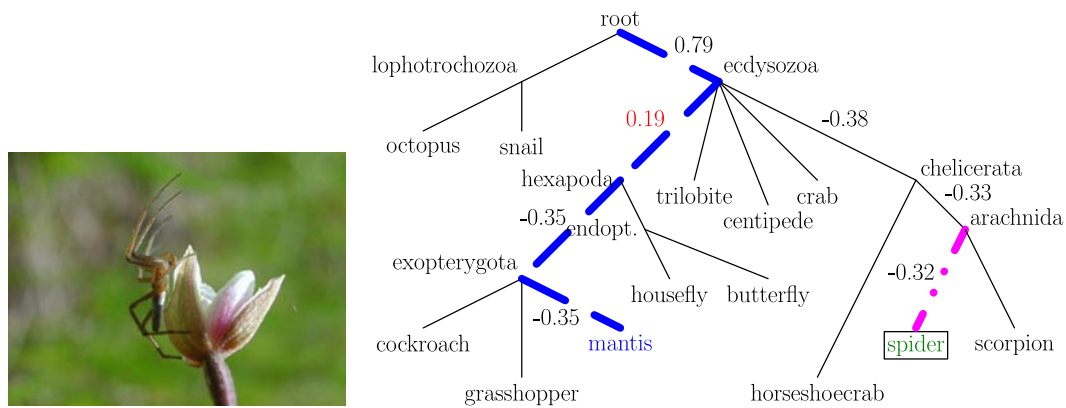
(upper) **hierarchic:** praying-mantis; **one versus all:** spider; **ground truth:** spider; Strong false positive vote for Hexapoda in hierarchical approach, the appearance of the spider does not show 8 legs clearly and is somewhat similar to mantids in pose and color.
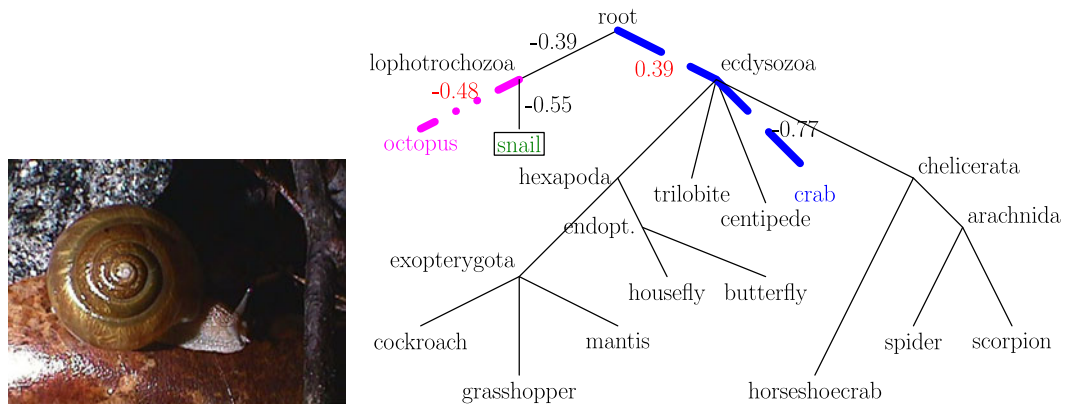
(lower) **hierarchic:** crab; **one versus all:** octopus; **ground truth:** snail; Strong false positive vote for Ecdysozoa causes hierarchy classifier to fail while one vs all predicts a taxonomically closer animal to the ground truth.

Fig. 7 (Color online) Example images where the hierarchical classifier is inferior to the one versus all baseline on Caltech animals, 13 classes. *Boxed green* denotes the ground truth label, *dashed blue* the path to the choice by hierarchical classifier and *dashdotted magenta* the decision by one versus all

votes at internal nodes which are too strong in order to be averaged out. Figure 8 shows cases when the hierarchical approach improves over a flat one versus all baseline. Typically the votes from internal nodes can average out and thus overrule false positive and too negative votes at the leaf nodes. The upper part of Fig. 8 shows a case when a taxonomically more plausible result can be achieved by using a hierarchy even when the classifier for the leaf node belonging the ground truth gives a too negative vote. In the lower part the hierarchic approach classifies the image correctly.

By comparing the confusion pattern of our taxonomy based procedure with that of the one-vs-all baseline, we observe clear qualitative differences. Figure 6 shows confusion differences between the two approaches (*y*-axis) versus the taxonomy losses (*x*-axis) for (a) bus and (b) cat of the VOC 2006 data. As expected, we can find the general tendency that the taxonomy based method confused more with the categories with lower taxonomy losses, while it can reduce the error with those with higher taxonomy

losses. We also checked significances of all confusion differences by a Wilcoxon signed-rank test from 20 random repetitions. Its p-values are summarized in the panel (c) (row: true classes, column: predicted classes). For instance, for (a) bus class, more images were correctly classified as bus (p-value = 0.06%) and confusion with person reduced significantly (0.16%) at the cost of increasing the error by prediction of cars (0.09%) which is in the taxonomy the closest category to bus. Similar relations hold for (b) cat class: confusions with the closer categories dog and horse increased, which brought improvements in confusions with farther away classes cow (0.4%), bicycle (3.1%) and motorbike (5.1%).

It is worth to point out that the improvement of taxonomy losses by hierarchical classification which was observed in Sect. 3 (see Table 3) and Sect. 4.5 implies that erroneous decisions are moved to lower levels in the hierarchy compared to baselines. This yields a more plausible, i.e. more human-like, result based on the taxonomy.
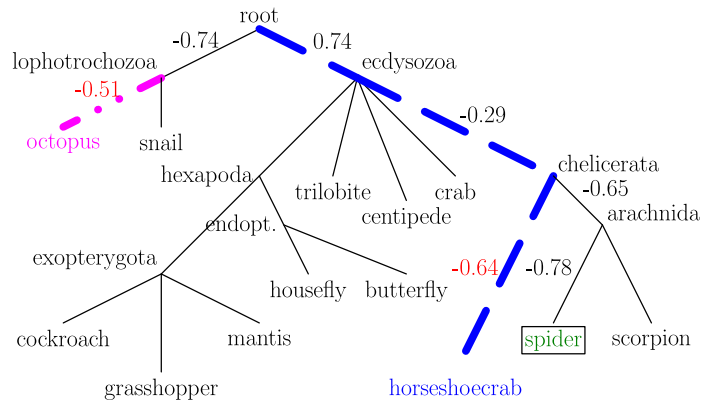
(upper) **hierarchic:** horseshoe crab; **one versus all:** octopus; **ground truth:** spider; The hierarchical approach predicts a horseshoe crab which belongs to the same subphylum Chelicerata as the spider, the score at the one vs all node for octopus is too large. The score in the one versus all node for horseshoe crab is too large, too, which prevents a correct classification as a spider.
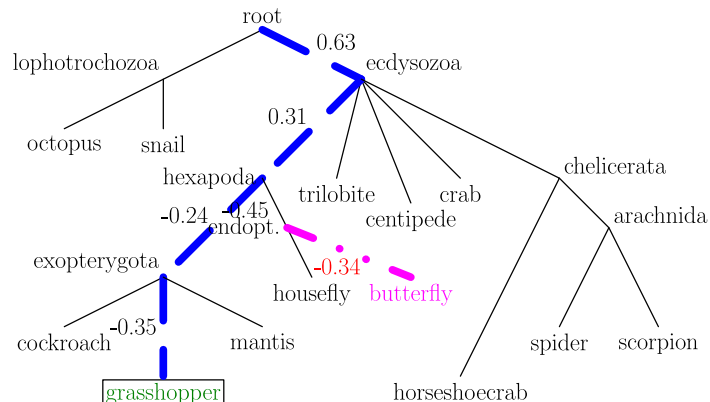
(lower) **hierarchic:** grasshopper; **one versus all:** butterfly; **ground truth:** grasshopper; The grasshopper gets classified correctly in the hierarchical approach at the Exopterygota versus all node which overrules the too low vote at the leaf nodes for class grasshopper compared to butterflies.

Fig. 8 (Color online) Example images where the hierarchical classifier outperforms the one versus all baseline on Caltech animals, 13 classes. *Boxed green* denotes the ground truth label, *dashed blue* the path to the choice by hierarchical classifier and *dashdotted magenta* the decision by one versus all

4.7.2 Comparison with Greedy Walks

We also analyzed the performance for local taxonomy approaches with hierarchical classification using greedy path-walks (Griffin and Perona 2008). We regard this direction rather as a side topic with respect to our comparison of structured versus local models. In this approach for each node in the taxonomy the set of negative examples is restricted to those with the class labels of the parent node. For example, for the class cat in the taxonomy from Fig. 2, a binary SVM is trained only with samples of classes Carnivora, i.e. cats and dogs. Such greedy walks lead to performance decrease. This is not surprising. Since the binary SVM at the leaf node 'cat' takes only images annotated with dog as negative samples, it may give highly positive scores to images containing horses or motorbikes. It is possible that the classifiers at the upper nodes, e.g. the nonlife-versus-life or the carnivora-versus-classifier misjudge some of these images and that the

cat-versus-dog classifier finally annotates them as cat with very high confidence.

We have found that the greedy walks strategy itself is detrimental. We obtain for both datasets a moderate rise in 0/1 loss and a sharp rise in taxonomy loss. In that sense the local approach adopted here is superior to other possible simpler local solutions. Performances of greedy walks can be found in Appendix.

The greedy approach has two advantages in running times compared to the local approach presented here. During training it deals at each node only with classifiers working on subsets of all categories which leads to a reduced amount of training data. During testing we have to follow only one path for each sample. The local approach presented here can be, in principle, modified by subsampling from the set of negative classes during training so that it uses the same amount of training data as the greedy approach. It would still retain the advantage of being able to suppress votes for outlier images as described above, i.e. when a car im-

Table 12 Mean AUCs on leaf nodes versus internal nodes for the local-SVM methods

Dataset	AUC leaf	AUC internal nodes
Caltech256 52 animals	88.49	84.82
Caltech256, 13 class subset	84.00	78.55
VOC2006 multi-class	86.38	91.40

age is tested in a cat versus dog classifier in a greedy walk scheme. While the greedy approach is the fastest option during test time, the local approach introduced here can be interpreted as a compromise between the structured SVMs and the greedy walks in terms of training and testing time. It achieves a trade-off between speed and precision.

4.7.3 AUC Scores at Leaf and Internal Nodes

We see from the real (Table 12) and synthetic datasets (see Sect. 3) a link between AUC scores at internal versus leaf nodes and the performance of flat zero one losses. When AUC scores are better on internal nodes as happened for VOC2006 and the synthetic data then the averaging in the hierarchical approach between them and the one versus all classifiers in the leaf nodes seems to improve flat zero one losses as well. This might be linked to alignment between visual similarities and taxonomy structure. With respect to taxonomy structure we note that half of the VOC2006 tree (see Fig. 2), namely the non life part, is constructed by subjective intuition of the authors and thus might be more similar to visual features. To give an example, the horse is part of odd-toed ungulates in a group with cats and dogs while the background appearance of horses, meadows, might be more similar to those of even-toed ungulates as cows and sheep. Thus this dataset might be more similar to our synthetic data where the visual features were perfectly aligned to the given taxonomy.

4.7.4 Generalization Ability for Learning of Superclasses in Taxonomies

The task of learning with taxonomies can be divided into two aspects. The first aspect is the optimization of a non flat loss via the taxonomy structure. We have seen in the preceding sections that taxonomy based methods do reduce taxonomy-induced losses.

The second aspect is that taxonomy based learning is an averaging with classifiers constructed by forming superclasses, in contrast for example to alternative learning approaches based on attributes (Farhadi et al. 2009; Lampert et al. 2009). We can conjecture in accordance to Fig. 1 that these superclasses can have larger variance in appearance compared to the single classes at the leafs. We may ask about the generalization capability of the used features

to superclasses used in internal nodes of the taxonomy. This aspect is in our opinion also linked to the question whether the usage of taxonomic models can reduce the flat zero one loss at all and how much the taxonomy loss can be reduced by taxonomic models.

Humans are able to generalize higher level categories very well, seemingly not worse or even better than more specific low level categories. For example humans can label cars very well even if their optical appearance is quite diverse as with old-timers, converted cars in strange shapes or rare car models, whereas identifying a specific car brand constitutes a more difficult task. This generalization capability seems to be uncommon for the current state of the art BoW feature extraction as we can see that false negative rates do increase considerably on intermediate nodes. We conjecture that classification with the BoW features suffers under the larger variability in appearance of high level concepts, which practically leads to a decreased SNR and an increase in nuisance dimensions. This might explain the differences between the ability of the taxonomy-based systems considered here and the assumed human performance.

One reason might be that we use in the classification scenario one bag of word feature which incorporates the base features from whole image in an unweighted manner. While this gives a state of the art in object classification competitions (Everingham et al. 2007, 2008, 2009) it might not be optimal for generalization to superclasses. Classical alternatives are part models (Ommer et al. 2006; Ommer and Buhmann 2010; Fergus et al. 2007) which are conceptually very appealing but do not seem to be widely used in competition systems for object recognition while being competitive in object localization scenarios (Dollár et al. 2008; Felzenszwalb et al. 2009). Their application is based on the assumption that the generalization to superclasses could be achieved by sharing parts. There exist potential but seemingly computationally more costly remedies in classification with good classification performances which learn a weighting in the space of base features of bag of word representations. Yang et al. (2008) cast classification a whole new learning problem directly in the space of base features and achieves very good scores for small sample sizes. Shahbaz Khan et al. (2009) achieve good classification performances by a weighting based on additional color features.

4.7.5 Outlook—Larger Numbers of Classes: Caltech256 Full

Here we consider the results for all 256 object classes from Caltech256. We omitted the clutter class and computed one k-means prototyped Bag of Words kernel based on 1000 words over the rgb color channel. We used 50 images per class and ten-fold crossvalidation which resulted in a training set size of 11520 samples. We were not able to compute

Table 13 Performance on Caltech256 all classes except for clutter, 10 splits

Method	Taxonomy loss	0/1 loss
One vs all	34.31 ± 0.74	**68.93** ± 1.23
Local tax AM	33.04 ± 0.7	72.91 ± 1.16
Local tax scaled GM	**32.77** ± 0.6	72.55 ± 1.14
Local tax greedy path-walk	37.81 ± 0.71	77.96 ± 1.3

the solutions from structured prediction methods however we are still able to compare one versus all against our local SVM approach. We observe in Table 13 qualitatively the same results as for the other, smaller, datasets. The taxonomy based approach improves on the taxonomy loss at the cost of setbacks in the zero one loss when compared to one versus all. The one versus all baseline performance ranges between the baseline used in Marszalek and Schmid (2008) and the best kernel from Gehler and Nowozin (2009).

5 Experiments on Real World Multilabel Datasets

Clearly the local SVM approach can also be used in a multilabel setting. Here in each image each concept can be present or absent independent of all other concepts.

We would like to emphasize that the target function evaluated here differs form the multiclass case as confusions between concepts are not well defined anymore. The idea of averaging classifiers is used here to enforce for each concept separately an ordering of images such that images of the concept in question and taxonomically close concepts are ranked highest.

Technically we replace confusion matrix based losses by threshold-independent ranking losses. A standard flat loss function used in the Pascal VOC challenge is Average Precision (AP) (Kishida 2005) and its mean over all classes. We assume that the pairs of SVM outputs and ground truth labels (z, y) are sorted according to the descending order of their output scores z_k over the sample index k. The average precision score is defined as

$$AP((z_k, y_k)_{k=1}^n) := \frac{1}{n} \sum_{i=1}^n \frac{1}{i} \sum_{k=1}^i \mathbb{I}\{y_k = 1\}. \tag{13}$$

The AP score is maximized when the images of the class in question are ranked first. It is invariant against permutation of the ordering of images from all other classes as long as the ranks of images from the class in question are untouched. However given relations from a taxonomy, we would prefer a ranking where images from taxonomically close classes are ranked in front of images from taxonomically far classes. To measure this difference we will use a second score, the Atax score.

The Atax score can be defined by replacing the hierarchy-unaware precision score in the AP measure by one minus the taxonomy loss to the taxonomy-nearest present class in the ground truth of the image and serves as a measure for binary problems which incorporates taxonomy information. For the multilabel taxonomy extension we consider instead of one binary label y_k a set along each class $\{y_k^{(r)} \in \{0, 1\}, r \in \{1, \dots, C\}\}$. Since we know for which class c of the multilabel problem we measure we can replace in the original AP score the 0/1-loss-based precision measurement by the minimal taxonomy distance δ between the measured class c and all positive labels in the ground truth $\{y_k^{(r)}, r \in \{1, \dots, C\} \mid y_k^{(r)} = 1\}$.

Again, we assume that the ground truth labels for class r, $y_k^{(r)}$ are sorted according to the descending order of the SVM outputs $z_k^{(c)}$ for class c.

$$ATax^{(c)}((z_k^{(c)}, \{y_k^{(i)}, i = 1, \dots, C\})_{k=1}^n)$$

$$:= \frac{1}{n} \sum_{i=1}^n \frac{1}{i} \sum_{k=1}^i 1 - \min_{r \in \{1, \dots, C\} \mid y_k^{(r)} = 1} \delta_T(c, r). \tag{14}$$

Since the taxonomy distance δ_T from (2) is scaled to lie in [0, 1] and a correct prediction leads to scores of $y_k = 1$ respectively $(1 - \min_{r \in \{1, \dots, C\} \mid y_k^{(r)} = 1} \delta(c, r)) = 1$, the ATax score is never smaller than the AP score. The precision used in AP scores can be interpreted as a zero-one discretization of the taxonomy score $(1 - \min_{r \in \{1, \dots, C\} \mid y_k^{(r)} = 1} \delta(c, r))$. Both scores have the advantage of being invariant against the classification threshold and evaluate the ranking of images. We do not use the ranking based scores for the multiclass problem. Inspecting the constraints of the structured prediction formulation from (6) shows that it aims at classifying each image correctly in the sense of obtaining a correct *ranking of classes* for each image. Its optimization does not aim at obtaining a correct *ranking of images* for each class. Thus, using a ranking score would be a biased measure against the structured approaches.

Note that for multilabel data the structured algorithms cannot be applied in their current form as the multi-class constraints are not well-defined anymore. Therefore we will compare one versus all versus local hierarchical approaches. As this frees us of time and memory consumption problems related to the structured algorithms we will use 20-fold crossvalidation. We will use the same features and kernels as described in Sects. 4.2 and 4.3 and measure with AP and ATax scores.

5.1 Datasets

VOC2006 multi-label data We use the VOC2006 dataset (Everingham et al. 2006) consisting of 10 object classes and 5301 images with its original, unmodified labels. The full taxonomy is given in Fig. 2.

Table 14 Performance on VOC06 as multilabel problem, 20 fold crossvalidation

Method	ATax	AP
One versus all	90.10 ± 3.46	80.13 ± 7.21
Local tax. scaled geometric mean	**91.29** ± 3.34	79.96 ± 7.23
Local tax. scaled, harmonic mean	90.85 ± 3.28	**80.61** ± 7.06

Table 15 Performance on VOC09 as multilabel problem, 20 fold crossvalidation

Method	ATax	AP
One versus all	79.02 ± 8.72	55.92 ± 15.91
Local tax. scaled geometric mean	**80.68** ± 8.20	54.62 ± 16.08
Local tax. scaled, harmonic mean	80.03 ± 8.33	**56.43** ± 15.77

VOC2009 multi-label classification task data This dataset consists of 20 classes with 7054 labeled images. It serves as a second multilabel setting for the local algorithms. The full taxonomy is given in Fig. 11 in the Appendix.

5.2 Experimental Results

Tables 14 and 15 show that even for a multilabel setting, introducing a taxonomy can improve taxonomy based as well as flat ranking scores, despite we have no notion of avoiding confusions anymore.

This may become relevant when using classifier scores for ranking images for retrieval. A higher ATax score implies that the desired class and similar classes are ranked higher than more distant classes which in effect leads to a subjectively improved ranking result from a human viewpoint. When looking for cats humans tend to be more impressed by results which return erroneously other pets than cars. Highly ranked images from very distant categories tend to be perceived as strong outliers.

Figure 9 shows examples where the hierarchical classifier is able to improve rankings simultaneously for classes which are far apart in the taxonomy given in Fig. 2. This shows that taxonomy learning for multilabel problems does not lead necessarily to mutual exclusion of taxonomy branches. In both images, the classes under consideration are separated already at the top level. We observe that images can be reranked to top positions despite average rankings at all nodes. For the upper image this occurs for the cow class, for the lower image this occurs for the motorbike class as can be seen from the rankings given along the paths. This can be explained by the property of the nonpositive p-means to be upper-bounded by the smallest score (see Sect. 2.4). Many images which achieved higher scores and ranks at some nodes along the considered path were effectively ranked lower because they received very low scores at least one

Table 16 Scaling of scores is important for multilabel problems, 20 fold crossvalidation

Method: local tax. arith. mean	ATax	AP
VOC06, *unscaled*	84.59 ± 6.73	60.31 ± 15.08
VOC06, *scaled*	89.58 ± 3.89	74.85 ± 8.51
VOC09, *unscaled*	73.35 ± 9.40	35.87 ± 14.73
VOC09, *scaled*	77.30 ± 9.45	46.58 ± 16.61

node in the same path. Note that the observed improvement in ranking is independent of the ranking loss.

Table 16 compares for both multilabel problems the performance of scaled versus unscaled combinations of scores. We see clearly that scaling of scores onto a compact interval contributes to the good performance of the local models. The good performance of scaled scores is not surprising as one can expect the SVM outputs to have different distribution statistics like variances across the nodes. Please note that for one versus all classification the scaling has no influence on the ranking scores as it is monotonous and rank-preserving and the score computation is done for each class separately.

6 Conclusions

When classifying complex data such as objects, humans are first of all much better than learning machines and most importantly human and machine errors diverge considerably. Among others, a reason for both findings is the impressing ability of humans to generate abstract representations that implicitly organize hierarchical knowledge and thus to create appropriate task relevant factorizations of the environment, put in one word humans generalize. One aspect of such abstract representation can be captured by taxonomies.

In this paper we have demonstrated that taxonomy-based learning using structured SVMs and local-SVM-based approaches on real world data yields improved results when measured with taxonomy-based losses. Local algorithms with generalized means voting perform on par to structured models while being considerably faster in training. The geometric mean appears to be a good a priori choice as a sensitivity tradeoff against small and large outliers. Successful minimization of taxonomy losses implies the reduction of confusions between distant categories, i.e. a step towards more human-like decision making. Note, however, that an improved result measured with taxonomy-based losses does not necessarily translate into a better result in a flat loss such as 0/1-loss since more meaningful confusions, i.e. improved quality of decision making does not necessarily come with overall quantitative improvements as other more meaningful confusions may come in addition—as a side effect. In the local SVM framework this can be checked by the AUC scores on the internal nodes compared to the leaf nodes.

Fig. 9 Example images where the hierarchical classifier improves rankings for taxonomically distant classes compared the one versus all baseline on VOC2006 multilabel problem. (*Upper*) car from 216 to 133, cow from 197 to 31. (*Lower*) motorbike from 108 to 52, person from 125 to 38
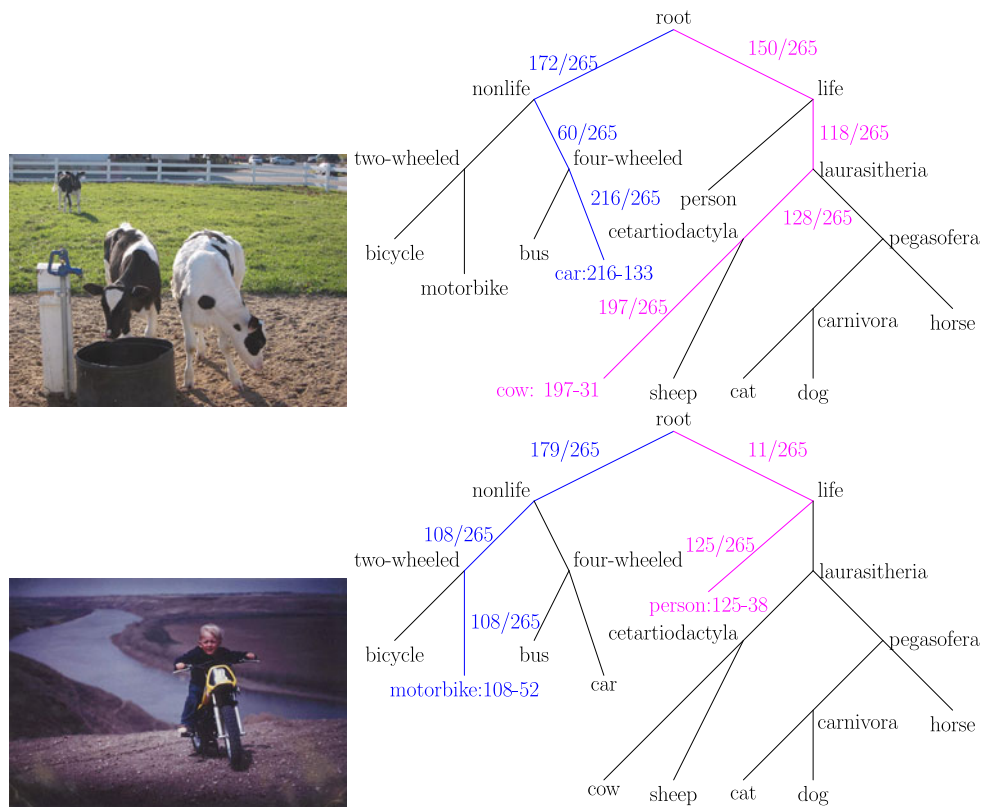
Experiments on synthetic data show, somewhat expectedly, that taxonomy based algorithms work better than the taxonomy-free baseline, when the data is aligned to the taxonomy. They suggest that performance gains are achieved for local procedures by combining classifiers with different trade-offs of false positive versus false negative rates. Interestingly but in fact to be expected, taxonomy based learners tend to make their errors rather close to the leaf-nodes of the taxonomy tree thereby confusing 'close' categories, whereas learners based on flat losses incur classification errors uniformly across the tree. The latter behaviour is one of the reasons to consider the decisions of flat loss trained learning machines more non-human than their taxonomy based counterparts.

The local as well as structured approaches can be combined with methods which learn taxonomies. The difference to previous approaches would be to measure taxonomy based errors instead of flat losses and rely in case of local algorithms on vote combination instead of reduced kernels and greedy path-walks. It is open in such case how much of the interpretation of a taxonomy is retained as a weak prior knowledge to define loss functions which penalize dissimilarities as they are perceived by humans.

A further direction was to compare the local-SVM procedures versus taxonomy-free multi-task learning approaches on multi-label problems. In these problems we are interested to rank the set of images for each class which demands

for threshold-invariant measures like the average precision scores for comparison or the Atax score. Our simulation study on VOC 2006 and 2009 shows encouraging results.

An overall challenge of the field would be to further the generic understanding of the different decision making between human and learning machine, ultimately combining low level machine precision, attribute based features and human abstraction optimally towards a truly cognitive automated decision making machinery.

Appendix

A.1 Detailed Experimental Results

The full comparison for Caltech256 animals 13 class subset and VOC2006 is shown in Tables 18 and 19.

Table 17 Performance on Caltech256 52 animals classes, 20 splits

Method	Taxonomy loss	0/1 loss
One vs all	30.66 ± 0.46	$\mathbf{62.56} \pm 0.67$
Struct mc mr $\Delta = \delta_T$	32.29 ± 0.35	66.91 ± 0.64
Struct mc sr $\Delta = \delta_T$	33.48 ± 0.39	68.86 ± 0.60
Struct mc sr $\Delta = \delta_{0/1}$	34.09 ± 0.38	68.05 ± 0.64
Local tax AM	30.01 ± 0.31	79.82 ± 0.55
Local tax scaled GM	$\mathbf{29.62} \pm 0.34$	76.19 ± 0.57
Local tax greedy path-walk	40.31 ± 0.34	77.65 ± 0.46
Struct tax mr $\Delta = \delta_T$	30.58 ± 0.31	81.19 ± 0.53
Struct tax sr $\Delta = \delta_T$	$-$[a] $\pm -$	$- \pm -$
struct tax sr $\Delta = \delta_{0/1}$	39.16 ± 0.45	76.85 ± 0.59

[a]Did not terminate after over seven days. Jobs consume over 20 GB

Table 18 Performance on Caltech256 animals 13 class subset data, 20 splits

Method	Taxonomy loss	0/1 loss
One vs all	42.49 ± 1.46	$\mathbf{57.04} \pm 1.98$
Struct mc mr $\Delta = \delta_T$	42.76 ± 0.96	64.35 ± 1.40
Struct mc sr $\Delta = \delta_T$	42.49 ± 1.49	57.06 ± 2.01
Struct mc sr $\Delta = \delta_{0/1}$	42.40 ± 1.29	57.05 ± 1.77
Local tax AM	41.78 ± 1.16	62.57 ± 1.42
Local tax scaled GM	$\mathbf{40.58} \pm 1.15$	58.33 ± 1.50
Local tax greedy path-walk	47.65 ± 1.13	63.33 ± 1.57
Struct tax mr $\Delta = \delta_T$	41.48 ± 1.22	61.54 ± 1.55
Struct tax sr $\Delta = \delta_T$	41.55 ± 1.65	58.21 ± 2.20
Struct tax sr $\Delta = \delta_{0/1}$	44.32 ± 1.07	59.22 ± 1.51

Table 19 Performance on VOC2006 as multi-class problem, 20 splits

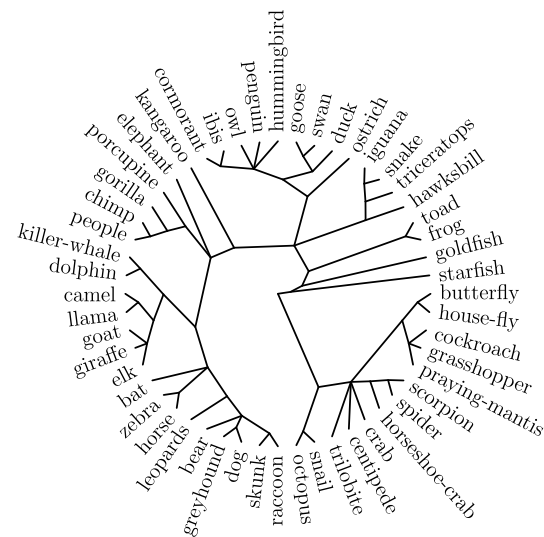| Method	Taxonomy loss	0/1 loss
One vs all	27.09 ± 1.88	50.54 ± 2.51
Struct mc mr $\Delta = \delta_T$	26.37 ± 1.77	51.04 ± 2.53
Struct mc sr $\Delta = \delta_T$	27.20 ± 1.89	50.73 ± 2.54
Struct mc sr $\Delta = \delta_{0/1}$	27.18 ± 1.87	50.70 ± 2.41
Local tax AM	26.02 ± 1.66	50.48 ± 2.34
Local tax scaled GM	25.86 ± 1.56	$\mathbf{50.10} \pm 2.29$
Local tax greedy path-walk	27.15 ± 1.65	51.85 ± 2.28
Struct tax mr $\Delta = \delta_T$	$\mathbf{25.78} \pm 1.67$	50.17 ± 2.17
Struct tax sr $\Delta = \delta_T$	27.24 ± 1.61	52.55 ± 2.23
Struct tax sr $\Delta = \delta_{0/1}$	27.63 ± 1.71	51.73 ± 2.50

Fig. 10 Taxonomy on 52 Animals Classes from Caltech256, the 13 class subset taxonomy is contained in the lower left quadrant from octopus to butterfly
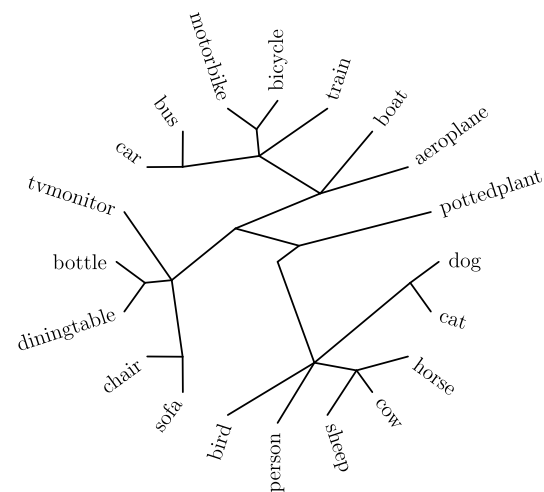

Fig. 11 Taxonomy on 20 Classes from Pascal VOC2009

References

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Blaschko, M. B., & Gretton, A. (2009). Learning taxonomies by dependence maximization. In *Advances in neural information processing systems*.

Bosch, A. (2007). Image classification for a large number of object categories. Ph.D. thesis, University of Girona.

Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proceedings of the conference on information and knowledge management*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. In *Machine Learning* (pp. 273–297).

Dollár, P., Babenko, B., Belongie, S. J., Perona, P., & Tu, Z. (2008). Multiple component learning for object detection. In *ECCV* (pp. 211–224).

Everingham, M., Zisserman, A., Williams, C. K. I., & Van Gool, L. (2006). The PASCAL visual object classes challenge 2006 (VOC2006) results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2007). The PASCAL visual object classes challenge 2007 (VOC2007) results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2008). The PASCAL visual object classes challenge 2008 (voc2008) results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2009). The PASCAL visual object classes challenge 2009 (voc2009) results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

Fan, X. (2005). Efficient multiclass object detection by a hierarchy of classifiers. In *CVPR* (pp. 716–723).

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. A. (2009). Describing objects by their attributes. In *CVPR* (pp. 1778–1785).

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2009). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **99**(1).

Fergus, R., Perona, P., & Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, *71*(3), 273–303.

Gehler, P., & Nowozin, S. (2009). On feature combination for multiclass object classification. In *ICCV*.

Griffin, G., & Perona, P. (2008). Learning and using taxonomies for fast visual categorization. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset* (Technical Report 7694). California Institute of Technology.

Har-Peled, S., Roth, D., & Zimak, D. (2002). Constraint classification for multi–class classification and ranking. In *Advances in neural information processing systems*.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods—support vector learning*. Cambridge: MIT Press.

Kishida, K. (2005). *Property of average precision and its generalization: an examination of evaluation indicator for information retrieval experiments* (Technical report). National Institute of Informatics, Japan.

Lafferty, J., Zhu, X., & Liu, Y. (2004). Kernel conditional random fields: representation and clique selection. In *Proceedings of the international conference on machine learning*.

Lampert, C. H., & Blaschko, M. B. (2008). A multiple kernel learning approach to joint multi-class object detection. In *Proceedings of the 30th DAGM symposium on pattern recognition*.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *CVPR* (pp. 951–958).

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE computer society conference on computer vision and pattern recognition* (Vol. 2, pp. 2169–2178). New York, USA.

Lowe, D. (2004). Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Marszalek, M., & Schmid, C. (2007). Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Marszalek, M., & Schmid, C. (2008). Constructing category hierarchies for visual recognition. In *Proceedings of the European conference on computer vision*.

Moosmann, F., Nowak, E., & Jurie, F. (2008). Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(9), 1632–1646.

Müller, K. R., Mika, S., Rätsch, G., Tsuda, S., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, *12*(2), 181–202.

Ommer, B., & Buhmann, J. M. (2010). Learning the compositional nature of visual object categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 501–516.

Ommer, B., Sauter, M., & Buhmann, J. M. (2006). Learning top-down grouping of compositional hierarchies for recognition. In *CVPRW'06: proceedings of the 2006 conference on computer vision and pattern recognition workshop* (p. 194), Washington, DC, USA. Los Alamitos: IEEE Comput. Soc.

Platt, J. (1999). In *Probabilistic outputs for support vector machine and comparison to regularized likelihood methods*.

Qi, G. J., Hur, X. S., & Zhang, H. J. (2009). Learning semantic distance from community-tagged media collection. In *MM'09: proceedings of the seventeen ACM international conference on Multimedia* (pp. 243–252).

Schölkopf, B., & Smola, A. J. (2001). *Learning with Kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning*. Cambridge: MIT Press.

Shahbaz Khan, F., van de Weijer, J., & Vanrell, M. (2009). Top-down color attention for object recognition. In *IEEE conference on computer vision (ICCV'09)*.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380.

Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., de Bona, F., Binder, A., Gehl, C., & Franc, V. (2010). The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, *11*, 1799–1802.

Tahir, M., van de Sande, K., Uijlings, J., Yan, F., Li, X., Mikolajczyk, K., Kittler, J., Gevers, T., & Smeulders, A. (2008). SurreyUVA SRKDA method. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/tahir.pdf.

Taskar, B., Guestrin, C., & Koller, D. (2004). Max–margin Markov networks. In *Advances in neural information processing systems*.

Tibshirani, R., & Hastie, T. (2007). Margin trees for high-dimensional classification. *JMLR*, *8*, 637–652.

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, *6*, 1453–1484.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1582–1596. http://doi.ieeecomputersociety.org/10.1109/TPAMI.2009.154.

Weston, J., & Watkins, C. (1999). Support vector machines for multiclass pattern recognition. In *ESANN* (pp. 219–224).

Yang, L., Jin, R., Sukthankar, R., & Jurie, F. (2008). Unifying discriminative visual codebook generation with classifier training for object category recognition. In *Proceedings of IEEE conference on computer vision and pattern recognition, IEEE* (pp. 1–8).

Zhang, J., Marszalek, M., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, *73*(2), 213–238.

Zweig, A., & Weinshall, D. (2007). Exploiting object hierarchy: combining models from different category levels. In *ICCV* (pp. 1–8).