Efficient Classification of Images with Taxonomies

Alexander Binder[1], Motoaki Kawanabe[1,2], and Ulf Brefeld[2]

[1] Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany
{alexander.binder, motoaki.kawanabe}@first.fraunhofer.de
[2] TU Berlin, Franklinstr. 28/29, 10587 Berlin, Germany
brefeld@cs.tu-berlin.de

Abstract. We study the problem of classifying images into a given, pre-determined taxonomy. The task can be elegantly translated into the structured learning framework. Structured learning, however, is known for its memory consuming and slow training processes. The contribution of our paper is twofold: Firstly, we propose an efficient decomposition of the structured learning approach into an equivalent ensemble of local support vector machines (SVMs) which can be trained with standard techniques. Secondly, we combine the local SVMs to a global model by re-incorporating the taxonomy into the training process. Our empirical results on Caltech256 and VOC2006 data show that our local-global SVM effectively exploits the structure of the taxonomy and outperforms multi-class classification approaches.

1 Introduction

Recognizing objects in images is one of the most challenging problems in computer vision. Although much progress has been made during the last decades, performances of state-of-the-art computer vision systems are far from the recognition rates of humans.

There are of course many natural explanations why humans outperform artificial recognition systems. However, an important difference between them is that humans effectively use background knowledge and incorporate semantic information into their decision making; their underlying representation is highly structured and allows for assessing co-occurrences to estimate the likeliness of events. By contrast, artificial recognition systems frequently rely on shallow or flat representations and models. The number of object recognition systems exploiting those co-occurrences or semantic relations between classes is rather small.

We believe that incorporating semantics into the object recognition process is crucial for achieving high classification rates. In this paper, we focus on tasks where the semantics is given *a priori* in form of a class-hierarchy or taxonomy. In general, incorporating a taxonomy into the learning process has two main advantages: Firstly, the amount of extra information that is added to the system details inter-class similarities and dependencies which can enhance the detection

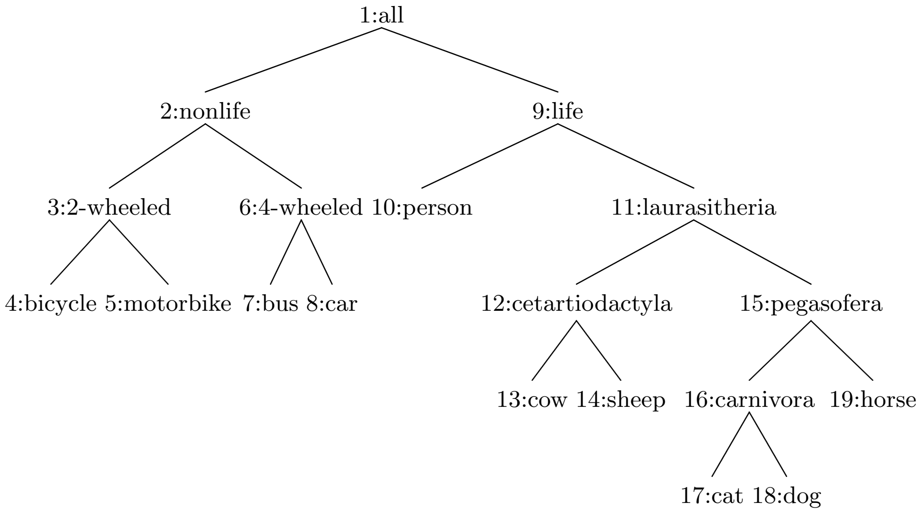

Fig. 1. The VOC2006 taxonomy.

performance. Secondly, the complexity of the task is spread across the taxonomy which can be exploited by simpler learning techniques.

There have been many publications dealing with *learning* class-hierarchies, for instance on the basis of delayed decisions [1], dependency graphs and co-occurrences [2, 3], greedy margin-trees [4], and by incorporating additional information [5]. By contrast, we focus on classifying images into a *pre-determined* taxonomy. The task fits into the structural learning framework [6, 7] which has recently gained much attention in the machine learning community and which has already been successfully applied to document classification with taxonomies [8].

However, the structural framework is computationally costly in terms of training time and memory consumption. We propose an efficient decomposition of the structural objective into several binary optimization tasks. The local models can be trained efficiently in parallel and converge to the same solution as their structural analogon. We furthermore show how to incorporate global taxonomy information into the training process of the local models by re-scaling the impact of images according to their location in the class-hierarchy. Empirically, we show on VOC2006 and Caltech256 data sets that our local-global SVM effectively exploits the structure of the taxonomy and outperforms multi-class classification approaches.

The remainder of this paper is structured as follows. Section 2 introduces the formal problem setting and Section 3 briefly reviews structural learning. We present our main theorem detailing the decomposition of the structured approach into local models in Section 4 where we also address the problem of assembling local models on a global level. We report on empirical results in Section 5 and Section 6 concludes.

2 Problem Setting

We focus on the following problem setting where we are given 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 $k(x, x') = \langle \phi(x), \phi(x') \rangle$. The set of labels is denoted by $Y = \{c_1, c_2, \ldots, c_k\}$. For simplicity, we focus on multi-class classification tasks, where every image is annotated by an element of Y. However, our approach can easily be generalized to the multi-label setting, where an image can be annotated with several class labels.

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 Figure 1 for an example. We assume the existence of a unique root node. The set of nodes on the path from the root node to a leave 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

$$\kappa_j(y) = \begin{cases} 1 : v_j \in \pi(y) \\ 0 : otherwise \end{cases} \quad 1 \leq j \leq |V|, \; y \in Y$$

such that the category *sheep* in Figure 1 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. As in the classical classification setting, we address this problem by searching a minimizer of the empirical risk that is defined on a fixed *iid* sample from P

$$R_{emp}(f) = \sum_{i=1}^{n} \delta\left(y^{(i)}, f(x^{(i)})\right). \tag{1}$$

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

$$\delta_{0/1}(y, \hat{y}) = \begin{cases} 0 & : \quad y = \hat{y} \\ 1 & : \quad \text{otherwise.} \end{cases} \tag{2}$$

When learning with taxonomies, the distance of y and $\hat{y}$ with respect to the taxonomy is fundamental. For instance, confusing a *bus* with a *cat* is more severe than mixing-up the classes *cat* and *dog*. We'll therefore also utilize a taxonomy-based loss function reflecting this intuition by counting the number of nodes between the true class y and the prediction $\hat{y}$,

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

For instance, the taxonomy-based loss between categories *horse* and *cow* in Figure 1 is $\delta_T(\mathrm{horse}, \mathrm{cow}) = 4$ because

$$\pi(\mathrm{cow}) \text{ xor } \pi(\mathrm{horse}) = \{\mathrm{cow}, \mathrm{cetartiodactyla}, \mathrm{pegasofera}, \mathrm{horse}\}.$$

3 Learning in Joint Input-Output Spaces

The taxonomy-based learning task matches the criteria for learning in joint input-output spaces [6, 7] where one learns a function

$$f(x) = \operatorname*{argmax}_y \langle w, \Psi(x, y) \rangle \tag{4}$$

that is defined jointly on inputs and outputs. The mapping $\Psi(x, y)$ is often called the joint feature representation and for learning taxonomies given by the tensor product [8]

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

Thus, the joint feature representation subsumes the structural information and explicitly encodes paths in the taxonomy. To minimize the empirical risk in Equation (1), parameters w can be optimized with conditional random fields (CRFs) [9] or structural support vector machines (SVMs) [6, 7]. Following the latter and using the formulation by [10, 11] we obtain the optimization problem in Equation (5).

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \sum_{\bar{y} \neq y^{(i)}} \xi_{\bar{y}}^{(i)} \\ \text{s.t.} \quad & \forall i,\ \forall \bar{y} \neq y^{(i)} : \quad \langle w, \Psi(x^{(i)}, y^{(i)}) - \Psi(x^{(i)}, \bar{y}) \rangle \geq \delta(y^{(i)}, \bar{y}) - \xi_{\bar{y}}^{(i)} \\ & \forall i,\ \forall \bar{y} \neq y^{(i)} : \quad \xi_{\bar{y}}^{(i)} \geq 0. \end{aligned} \tag{5}$$

The above minimization problem has one constraint for each alternative classification per image. Every constraint is associated with a slack-variable $\xi_{\bar{y}}^{(i)}$ that acts as an upper bound on the error δ caused by annotating the i-th image with label $\bar{y}$. Once, optimal parameters w^* have been found, these are used as plug-in estimates to compute predictions for new and unseen examples using Equation (4). The computation of the argmax can be performed by explicit enumeration of all paths in the taxonomy.

Note that the above formulation differs slightly from [6, 7] where every instance is associated with only a single slack variable representing the most strongly violated constraint for each image. Although, Equation (5) 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 infeasible. As a remedy, we will present an efficient decomposition of the structural optimization problem in the next section.

4 Local-Global Support Vector Learning

In this section we present the main contribution of this paper. Firstly, we devise a decomposition of the structural approach in Equation (5) into several local models in Section 4.1. Secondly, we show how to combine the local models globally by incorporating the structure of the taxonomy into the learning processes in Section 4.2.

4.1 An Efficient Local Decomposition

The idea is to learn a binary SVM 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 an intractable structural approach. To preserve the predictive power, the final binary SVMs need to be assembled appropriately according to the taxonomy. Essentially, our approach boils down to training $|V|$ independent binary support vector machines 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., $Pr(\kappa_j(y) = 1)$. It will be convenient to define the auxiliary label function $z_j(y)$ by

$$z_j(y) = \begin{cases} +1 : \text{if } \kappa_j(y) = 1 \\ -1 : \text{otherwise.} \end{cases} \tag{6}$$

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. In Figure 1 for instance, we have $z_{\text{life}}(\text{cow}) = 1$ but $z_{\text{life}}(\text{bus}) = -1$.

Using Equation (6), we resolve the *local-SVM* optimization problem that can be split into $|V|$ independent optimization problems, effectively implementing a one-vs-rest classifier for each node.

$$\begin{aligned} \min_{\tilde{w}_j, \tilde{b}_j, \tilde{\xi}_j} \quad & \frac{1}{2} \sum_{j=1}^{|V|} \|\tilde{w}_j\|^2 + \sum_{j=1}^{|V|} \tilde{C}_j \sum_{i=1}^{n} \tilde{\xi}_j^{(i)} \\ \text{s.t.} \quad & \forall i,\ \forall j : \quad z_j(y^{(i)})(\langle \tilde{w}_j, \phi(x^{(i)}) \rangle + \tilde{b}_j) \geq 1 - \tilde{\xi}_j^{(i)} \\ & \forall i,\ \forall j : \quad \tilde{\xi}_j^{(i)} \geq 0. \end{aligned} \tag{7}$$

At test time, the prediction for new and unseen examples is computed similarly to Equation (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 leave y,

$$\hat{y} = \operatorname*{argmax}_{y \in Y} \sum_{j:\kappa_j(y)=1} f_j(x). \tag{8}$$

The following theorem shows that the above approach is equivalent to the structural SVM in Equation 5.

Theorem 1. *If $C = \tilde{C}_j$ for $1 \leq j \leq |V|$ and $\delta(y, \bar{y})$ in Equation (5) is the 0/1 loss (Equation (2)) then the optimization problems in Equations (5) and (7) are equivalent.*

The proof is shown in the Appendix and relies on projecting combinatorial variables $\bar{y}$ onto nodes, hence reducing the number of possible events significantly to only a binary choice: either a node lies on a path or not. Along with the number of combinatorial outcomes, the training times reduce significantly. Another appealing aspect of this result is that the $|V|$ support vector machines can be trained efficiently in parallel. This property is also preserved when re-incorporating the taxonomy information as is shown in the next section. Moreover, model selection can be applied to the training process of each model separately which may lead to highly adapted local models with optimal trade-off C_j parameters (and potentially also kernel parameters) while its structural counterpart allows only for a single parameter C. In the next section we will show how to combine the local SVMs of optimization problem (7) globally by introducing example-specific costs.

4.2 Incorporating Global Misclassification Costs

The previous section shows how to decompose the structural approach into independent, binary problems. Although, the taxonomy is still necessary for scoring paths at prediction time (Equation (8)), the training processes of the binary SVMs is independent of any taxonomy information.

We now show how to incorporate taxonomy information into the training process of the local models. The intuition behind our approach is to reweight images by their taxonomy-distance. That is, we intend to penalize confusions of classes that have a large distance with respect to the taxonomy. On the other hand we are willing to accept misclassifying instances of nearby classes.

To be precise, we identify the cost $c_j(x^{(i)})$ at node v_j for a negative example as the number of nodes on the path from the j-th node to the true output; that is, $c_j(x^{(i)}) = \delta_T(v_j, y^{(i)})$. For instance, in Figure 1, the associated costs with an instance (x, bus) at the node *life* are $c_{\text{life}}(x) = 4$. The costs for positive examples are given by the costs of all negative instances for balancing reasons,

$$c_j(x) = \frac{1}{n_j^+} \sum_{i:z_j(y^{(i)})=-1} c_j(x^{(i)}),$$

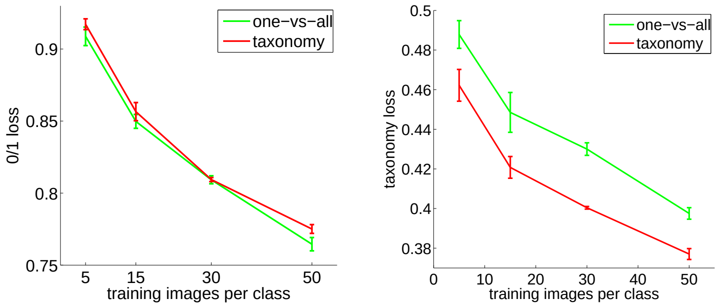

Fig. 2. Results for Caltech256. 0/1 loss and Taxonomy loss of local-global-SVM.

where n_j^+ is the number of positive examples at node v_j. Given the weights c_j, these can be augmented into the training process according to [12]. The *local-global* SVM optimization problem can be stated as follows,

$$\begin{aligned} \min_{\tilde{w}_j, \tilde{b}_j, \tilde{\xi}_j} \quad & \frac{1}{2} \sum_{j=1}^{|V|} \|\tilde{w}_j\|^2 + \sum_{j=1}^{|V|} \tilde{C}_j \sum_{i=1}^{n} c_j(x^{(i)})\, \tilde{\xi}_j^{(i)} \\ \text{s.t.} \quad & \forall i,\ \forall j: \quad z_j(y^{(i)})(\langle \tilde{w}_j, \phi(x^{(i)}) \rangle + \tilde{b}_j) \geq 1 - \tilde{\xi}_j^{(i)} \\ & \forall i,\ \forall j: \quad \tilde{\xi}_j^{(i)} \geq 0. \end{aligned} \tag{9}$$

That is, if $c_j(x^{(i)}) \gg 1$ then the importance of the i-th input is increased while $c_j(x^{(i)}) \ll 1$ decreases its impact on the objective function. Thus, input examples that are associated with large costs $c_j(x)$ are likely to be classified correctly while accepting misclassifications associated with small costs.

5 Empirical Results

We compare our local-global SVMs empirically with the one-vs-rest SVM which is contained as a special case of our approach and furthermore equivalent to employing a flat taxonomy, where the root is directly connected to all leave nodes.

We experiment on the Caltech256 [13] and on the VOC2006 [14] data sets.

5.1 Data Sets

The Caltech256 data set comes with 256 object categories plus a clutter class; we focus on the 52 animal classes. This reduces the number of images to 5895; the smallest class has 80, the largest 270 elements. Each image is annotated with precisely one class label. We construct 5 sets of training, holdout, and test splits and deploy a taxonomy with approximately 100 nodes from biological systematics as underlying class-hierarchy. The left panel of Figure 3 shows the loss $\delta_T(y, \hat{y})$ based on our taxonomy. Here blue color denotes categories which are close in taxonomy distance while red pairs are far apart. For example, the classes 40–52 belong to a sub-group which is far from the cluster 18-39.

The VOC2006 dataset comprises 5,304 images containing in total 9507 annotated objects from 10 categories. The smallest class consists of 354 and the largest contains 1341 examples. We prepare 5 different training, holdout, and test splits by drawing images randomly to preserve the same number of class labels as proposed by the VOC2006 challenge. Thus, our training sets vary in their sizes and comprise between 2,500 and 3,000 instances. Although VOC2006 is a multi-label task, we treat the data set as a multi-class classification task by comparing for each class and each image belonging to that class the class label to the class of the maximum score. The taxonomy is shown in Figure 1.

5.2 Feature Extraction and Combination

We employ pyramid histograms [15] of visual words [16] (PHOW) for pyramid levels 0,1 over grey, opponent color 1 and 2 channels, which results in six different features. For every color channel, 1200 visual words are computed by hierarchical k-means clustering on SIFT features [17] from randomly drawn images. For VOC2006, the underlying SIFT features are extracted from a dense grid of pitch six. For Caltech256 the images have been pre-scaled to have 160,000 pixels, while their aspect ratios have been preserved. We apply a χ^2-kernel for every PHOW feature [18]. The kernel width parameter is initialized with the mean of the χ^2 distances over the respective training splits [2]. The final kernel K is then computed by the product of the six χ^2-kernels, $K = \left(\prod_{i=1}^{6} K_i\right)^{\lambda}$, where λ controls the width of the product kernel.

5.3 Experimental Setup

Model selection is performed for the SVM trade-off parameter C in the range $C \in [6^{-2}, 6^4]$ and for the kernel parameter λ in the interval $\lambda \in [3^{-7}, 3^2]$. For experiments with the taxonomy loss δ_T (Equation (3)) we also apply δ_T for finding the optimal parameters in the model selection. All other experiments use the 0/1-loss analogon. We deploy class-wise losses at each node to balance extreme class ratios for all methods. In our binary classification setting, this reduces to the computing the average of the loss on the positive class $\ell(+1)$ and that of the negative class $\ell(-1)$. The final value is then given by $\ell = \frac{1}{2}(\ell(+1) + \ell(-1))$. We use the model described in Section 4.2 and refer to it as *local-global SVM.*

5.4 Caltech256

Figure 2 shows the results for varying numbers of training images per class for combining the training of local-global SVMs (right). As expected, the error of all methods decrease with the sample size. As expected, there is no significant

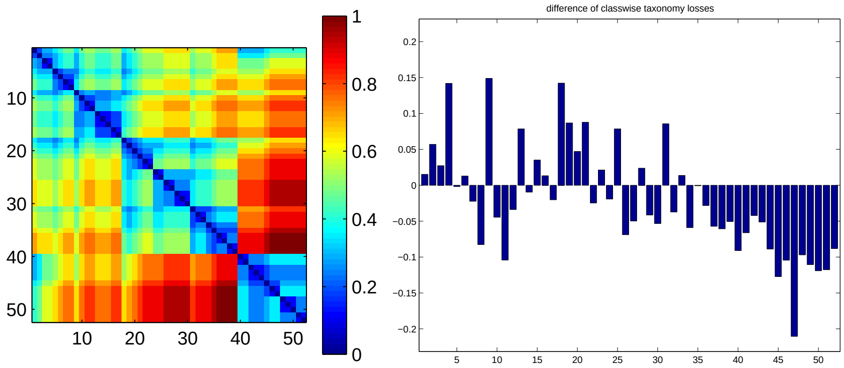

Fig. 3. (Left panel) The taxonomy loss $\delta_T(y, \hat{y})$ for the Caltech256 experiment. (Right panel) The expected taxonomy loss for each class.

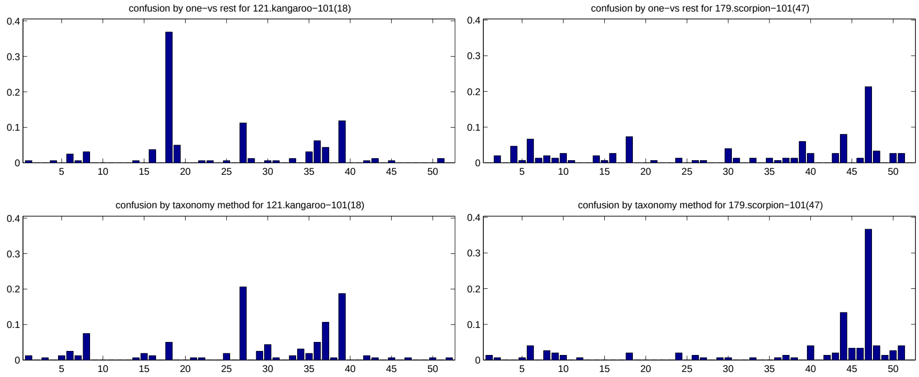

Fig. 4. Confusion probabilities for classes kangaroo (left) and scorpion (right).

difference between a one-vs-all SVM and our local-global SVM in terms of 0/1 loss. By contrast, the local-global SVM significantly outperforms the shallow basline in terms of taxonomy loss δ_T. This effect is due to incorporating the taxonomy structure into the training process of local-global SVMs.

To interpret this result, we compute average confusion matrices detailing $P(\hat{y}|y)$ over 5 repetitions for 50 training images per class. We compute the average taxonomy loss with respect to the confusion probabilities for each object class, i.e., $\sum_{\hat{y}} \delta_T(y, \hat{y}) P(\hat{y}|y)$. The right panel of Figure 3 shows the differences of the average taxonomy losses between our method and the one-vs-rest baseline. Negative values in this plot indicate that our method reduces the taxonomy loss of the corresponding classes. We observe that the local-global SVM effectively reduces the taxonomy loss for a large number of classes. However, there also exist classes such as *toad* (4), *ostrich* (9), and *kangaroo* (18) for which the error increased. To investigate this finding, we compared confusion probabilities of the baseline (upper left panel) and the taxonomy-based approach (lower left panel) for the *kangaroo* class in Figure 4. In fact, *kangaroo* was substantially confused

Table 1. Error-rates for VOC2006.

	δ_{01}	δ_T
one-vs-rest	0.5257 ± 0.0131	0.2714 ± 0.0050
taxonomy	0.5006 ± 0.0126	0.2507 ± 0.0042

with *llama* (27) and *raccoon* (39) which are rather far from *kangaroo* in our taxonomy.

By contrast, our approach achieves significantly better accuracies than the baseline on the *scorpion* (47) class. Figure 4 (top right panel) shows that the taxonomy model increases confusions when compared to one versus all slightly between scorpion and Arthropoda like *crab* (44) which are relocated in the higher fourty indices and are biologically close to scorpions while it reduces confusions for example to *kangaroo* (18), *raccoon* (39) and *toad* (4).

Our analysis indicates that a mismatch between the similarity in feature space and distance with respect to the taxonomy can substantially harm the classification performance. Thus to improve learning with pre-determined taxonomies, one would either have to (i) remove these mismatches by reverse engineering the class-hierarchy or to (ii) design features which resolve this conflict. We will address both aspects in future research.

5.5 VOC2006

Finally, Table 1 shows average precisions for the VOC2006 data set. The left column shows the 0/1 loss (Equation (2)) and the loss in the right column corresponds to the average number of nodes that lie in-between the true and the predicted class (Equation (3)). For both loss functions, the local-SVM yields significantly lower error-rates than a flat one-vs-rest classification.

6 Conclusions

We presented an efficient approach to classification of images with underlying taxonomies. Our method grounds on decomposing structural support vector machines into local, binary SVMs that can be trained in parallel. Furthermore, we employed taxonomy-based costs for images to incorporate the taxonomy into the learning process. Significant contributions like [1, 19] compared taxonomy models to flat ones using *0/1-loss*. Empirically, we observed our local-global SVMs to effectively benefit from the underlying taxonomy with respect to *taxonomy loss*: our approach was always equal or better than its shallow multi-class counterpart that cannot make use of taxonomy information.

Acknowledgements This work was supported in part by Federal Ministry of Economics and Technology of Germany under the project THESEUS (01MQ07018) and by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886.

References

1. Marszalek, M., Schmid, C.: Constructing category hierarchies for visual recognition. In: Proceedings of the European Conference on Computer Vision. (2008)
2. Lampert, C.H., Blaschko, M.B.: A multiple kernel learning approach to joint multiclass object detection. In: Proceedings of the 30th DAGM symposium on Pattern Recognition. (2008)
3. Blaschko, M.B., Gretton, A.: Learning taxonomies by dependence maximization. In: Advances in Neural Information Processing Systems. (2009)
4. Tibshirani, R., Hastie, T.: Margin trees for high-dimensional classification. JMLR **8** (2007) 637 – 652
5. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2007)
6. Taskar, B., Guestrin, C., Koller, D.: Max–margin Markov networks. In: Advances in Neural Information Processing Systems. (2004)
7. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of Machine Learning Research **6** (2005) 1453–1484
8. Cai, L., Hofmann, T.: Hierarchical document categorization with support vector machines. In: Proceedings of the Conference on Information and Knowledge Management. (2004)
9. Lafferty, J., Zhu, X., Liu, Y.: Kernel conditional random fields: representation and clique selection. In: Proceedings of the International Conference on Machine Learning. (2004)
10. Weston, J., Watkins, C.: Multi–class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Sciences, Royal Holloway, University of London (1998)
11. Har-Peled, S., Roth, D., Zimak, D.: Constraint classification for multi–class classification and ranking. In: Advances in Neural Information Processing Systems. (2002)
12. Brefeld, U., Geibel, P., Wysotzki, F.: Support vector machines with example dependent costs. In: Proceedings of the European Conference on Machine Learning. (2003)
13. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology (2007)
14. Everingham, M., Zisserman, A., Williams, C.K.I., Gool, L.V.: The 2006 pascal visual object classes challenge (voc2006) results (2006)
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2., New York, USA (June 2006) 2169–2178
16. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic (May 2004) 1–22
17. Lowe, D.: Distinctive image features from scale invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
18. Zhang, J., Marszalek, M., S.Lazebnik, Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73**(2) (2007) 213–238

19. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008)

Appendix: Proof of Theorem 1

Proof: We show the equivalence of the unconstraint objective functions. We first note that the dual representation of the structural parameter vector is given by $w = \sum_{i,\bar{y}\neq y_i} \alpha(i,\bar{y})(\Psi(x_i,y_i) - \Psi(x_i,\bar{y}))$. Since nodes are treated independently and the κ_j are orthogonal, we have

$$\begin{aligned}
\|w\|^2 &= \left\| \sum_{i=1}^{n} \sum_{\bar{y}\neq y^{(i)}} \alpha(i,\bar{y}) \left(\Psi(x^{(i)},y^{(i)}) - \Psi(x^{(i)},\bar{y}) \right) \right\|^2 \\
&= \sum_{j=1}^{|V|} \left\| \sum_{i=1}^{n} \sum_{\bar{y}\neq y^{(i)}} \alpha(i,\bar{y})\, \phi(x^{(i)}) \left(\kappa_j(y^{(i)}) - \kappa_j(\bar{y}) \right) \right\|^2 \\
&= \sum_{j=1}^{|V|} \left\| \sum_{i=1}^{n} \tilde{\alpha}_j(i)\, z_j(i)\, \phi(x^{(i)}) \right\|^2 \\
&= \sum_{j=1}^{|V|} \|w_j\|^2,
\end{aligned}$$

for $\tilde{\alpha}_j(i) = \sum_{\bar{y}\neq y^{(i)}} \alpha(i,\bar{y})|\kappa_j(y^{(i)}) - \kappa_j(\bar{y})|$. Note that the pseudo labels in Equation (6) can alternatively be computed by $z_j(i) = \mathrm{sign}(\sum_{\bar{y}\neq y^{(i)}} \kappa_j(y^{(i)}) - \kappa_j(\bar{y}))$. For the sum of the slack variables, we define the non-negativity function $(t)_+ = t$ if $t > 0$ and 0 otherwise and proceed as follows:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{\bar{y}\neq y^{(i)}} \xi_{\bar{y}}^{(i)} &= \sum_{i=1}^{n} \sum_{\bar{y}\neq y^{(i)}} \left(1 - \langle w, \Psi(x^{(i)},y^{(i)}) \rangle + \langle w, \Psi(x^{(i)},\bar{y}) \rangle \right)_+ \\
&= \sum_{j=1}^{|V|} \sum_{i=1}^{n} \sum_{\bar{y}\neq y^{(i)}} \left(1 - \langle w_j, \phi(x^{(i)}) \rangle \left[\kappa_j(y^{(i)}) - \kappa_j(\bar{y}) \right] \right)_+ \\
&= \sum_{j=1}^{|V|} \sum_{i=1}^{n} \left(1 - z_j(i) \langle \tilde{w}_j, \phi(x^{(i)}) \rangle \right)_+ \\
&= \sum_{j=1}^{|V|} \sum_{i=1}^{n} \tilde{\xi}_j^{(i)},
\end{aligned}$$

where w_j denotes the j-th block of $w = (w_1, \ldots, w_{|V|})$ and is given by

$$\tilde{w}_j = w_j | \sum_{i,\bar{y}\neq y^{(i)}} \kappa_j(y^{(i)}) - \kappa_j(\bar{y})|.$$

This concludes the proof. □