Learning with Few Examples Using a Constrained Gaussian Prior on Randomized Trees

Erik Rodner, Joachim Denzler

Chair for Computer Vision, Friedrich Schiller University of Jena
Email: {rodner,denzler}@informatik.uni-jena.de

Abstract

Machine learning with few training examples always leads to over-fitting problems, whereas human individuals are often able to recognize difficult object categories from only one single view. It is a common belief, that this is mostly established by transferring knowledge from related classes. Therefore, we introduce a *new hybrid classifier* for learning with very few examples by exploiting interclass relationships. The approach consists of a randomized decision trees structure which is significantly enhanced using maximum a posteriori (MAP) estimation.

For this reason, a constrained Gaussian is introduced as a *new parametric family* of prior distributions for multinomial distributions to represent shared knowledge of related categories. We show that the resulting MAP estimation leads to a simple recursive estimation technique, which is applicable beyond our hybrid classifier.

Experimental evaluation on two public datasets (including the very demanding Mammals database) shows the benefits of our approach compared to the base randomized trees classifier.

1 Motivation and Introduction

Learning to recognize objects of different categories is one of the major research fields within computer vision and machine learning. Although current state-of-the-art approaches reach impressive results on difficult datasets [3], they are not able to handle very small training sets.

Without additional information, machine learning with few training examples always reduces to an ill-posed optimization problem. It is a common paradigm within the community that information of similar object categories or classification tasks (*support classes / tasks*) is the most useful source to enhance generalization ability of weak representations [4]. This principle is known in the literature as learning to learn, knowledge transfer or transfer learning. Therefore handling few training examples needs classifiers that use interclass relationships (interclass transfer).

Research in this field is often motivated by closing the gap between human and machine vision quality. However, learning with weak representations is also a typical task demanded by the industry. Gathering training material is often expensive and time-consuming and can have a significant impact on the overall cost of resulting systems.

Previous work on interclass transfer differs in the type of information transferred. Miller et al. [17] try to estimate shared geometric transformations, which can be applied indirectly to a new class representation. Another idea is to assume shared structures in feature space and estimate a metric or transformation from support classes [7, 19, 1]. This mostly leads to methods similar to linear discriminant analysis, without clear motivation and suitable comparisons to the standard approach.

Torralba et al. [21] use a discriminative boosting technique which exploits shared class boundaries within feature space to allow more efficient multiclass learning with a sub-linear number of features. In contrast, Fei-Fei et al. [5] develop a generative framework with MAP estimation of model parameters using a prior distribution estimated from support classes. Contextual information as a helpful information source in object recognition systems can be regarded as a special case of interclass transfer [8, 12, 20].

Our work is motivated by the basic ideas in [21, 5] and combines them in a new manner. We propose to use a combined discriminative and generative technique as a framework of a new transfer learning approach. The key concept is a *MAP esti-*

mation of parameters of a discriminative classifier. In contrast to other hybrid methods, such as [13], feature space is efficiently partitioned using a discriminative random forest classifier (or extremely randomized trees (ERT), [10, 2]) . This partitioning and an additional prior on the distribution within each decision tree is afterwards used as prior knowledge in a MAP estimation of model parameters of a new class with few training examples (Figure 1).

This allows to develop a classifier, which efficiently combines the generalization power of discriminative- and the advantage of generative approaches to provide a better models of weak representations.

The use of ensembles of trees, which is the base discriminative part of our method, prevents overfitting by randomization and thus solves one of the main problems original decision tree classifiers suffer from. Recent applications [20, 6, 15] show the great potential and wide applicability of this base classifier.

We first review randomized decision trees as presented by [10]. Next we show that Bayesian estimation using a prior distribution is a well founded possibility to transfer knowledge from related classes and how to apply this idea to decision trees. In section 4 a compact model for multinomial distributions, called constrained Gaussian prior, is introduced, which is the theoretical key concept of our method. Finally, experiments on public available datasets demonstrate the benefits of our hybrid classifier. This includes an evaluation using the very difficult Mammals database of [9]. The use of this database leads to a short discussion (section 5.3) about the ability of our method to naturally transfer context. A summary of our findings conclude the paper.

2 Randomized Decision Trees

The discriminative part of our approach is based on extremely randomized trees (ERT) as introduced by Geurts et al [10]. Decision tree classifiers are (mostly) binary trees which store a weak classifier and posterior distributions $p(\Omega_i \mid n)$ for each class i at all nodes n. A weak classifier is a binary function consisting of an one-dimensional feature and a simple threshold. At classification phase an example (feature vector, image) is traversed down the tree according to the output of each weak classifier on the path until a leaf node is reached. The posterior of the leaf node is the result of the decision tree and estimates the probabilities of an arbitrary example of class i to reach the leaf.

Standard decision tree approaches suffer from two serious problems: long training time and overfitting. The ERT approach solves both issues by random sampling. Training time is easily reduced by an approximate search for the most informative weak classifier in each node, instead of evaluating every feature and threshold. The selection is done by choosing the weak classifier with the highest gain in information from a random fraction of features and thresholds.

Given enough training data for each class i, generalization performance can be greatly increased by learning S decision trees (forest) with a random subset of the training data. Given the leaf nodes of the forest $\mathbf{n}^L = (n_1^L, \ldots, n_S^L)$ the overall posterior can be obtained by simple averaging [2]:

$$p(\Omega_i \mid \mathbf{n}^L) = \frac{1}{S} \sum_{s=1}^{S} p(\Omega_i \mid n_s^L) \; . \qquad (1)$$

This special case of Bagging [2] reduces overfitting effects without the need of additional tree pruning. Unbalanced training sets lead to a biased decision tree classifier, therefore we follow [20] and weight each example with the inverse class frequency.

3 Transfer Learning using ERT

Learning with few or only one single example of each class often leads to over-fitting of the model parameters estimated by the classifier. When using decision trees, the ability of random forests to efficiently learn robust classifier combinations cannot be used due to the fact, that they need to additionally reduce the training data. Within each tree, a single example only increases posterior probabilities in a single branch, which can be seen as memorizing the training example completely. For this reason our approach uses interclass transfer to learn shared features from support classes and incorporates this information to boost the generalization ability of a new class. The main steps of our algorithm are summarized in Figure 2.

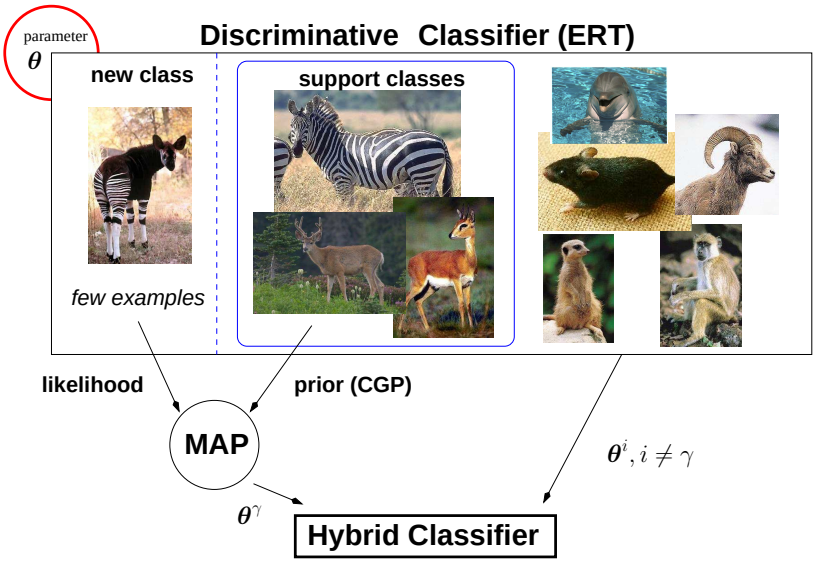

Figure 1: General concept of our approach for transfer learning in the context of learning with few examples: common knowledge about the recognition task is obtained from training data of related similar classes (support classes) and used to build a classifier which is able to efficiently learn from a weak representation (new class). Images are collected from the Mammals database of [9], which is used for experiments in section 5.

3.1 Discriminative Structure

Despite the well founded, probabilistic background of Bayesian methods, discriminative classifiers often lead to a better generalization performance, which can be seen in the success of support vector machines as a standard classifier within the machine learning community.

For this reason we propose to use as a first step a discriminative random forest trained with examples of all classes. All subsequent steps to incorporate knowledge of the new class, use this fixed discriminative structure. This concept has been used in [11, 15], to reuse features and reduce computation times. Our approach can be seen as a refinement step of this fixed discriminative classifier using a Bayesian framework. In the following, we describe our method for a single tree in the forest.

3.2 Generative Transfer Learning

As proposed by [5], transfer learning can be done by MAP estimation of model parameters $\boldsymbol{\theta}$ related to the new class γ. Knowledge from a set $\mathcal{S}$ of support classes is incorporated within this Bayesian framework as a prior distribution of parameters $\boldsymbol{\theta}$. These classes represent a subset of all available categories, which are assumed to be related to the new class. Therefore the fundamental assumption is, that it is possible to estimate a suitable prior distribution useful to regularize parameter estimation of a related class. This assumption can be expressed mathematically by :

$$\boldsymbol{\theta}^{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(T^\gamma \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid T^{\mathcal{S}}) \,. \quad (2)$$

where T^γ denotes training data of the new class and $T^{\mathcal{S}}$ denotes training data of all support classes. Please note that this directly incorporates the main assumption by using a prior distribution depending on the training data of all support classes $T^{\mathcal{S}}$.

With a fixed tree structure, incorporating training data of the new class reduces to estimating posterior distributions in the leafs. Using a single training example and updating posterior distributions without additional knowledge induces the update of only a

(I) Learn a randomized decision forest using all classes [10].
(II) Calculate node probabilities for each node and each class (equation (3)).
(III) Approximate prior knowledge using a constrained Gaussian prior (equation (6)).
(IV) MAP estimation of leaf probabilities of the class with few training examples using a CGP (section 4.1).
(V) Calculate node posteriors from estimated leaf probabilities.
(VI) Build additional discriminative layers (section 3.4).

Figure 2: Main steps of our approach. Steps II to VI are performed independently for each tree of the forest.

single branch from the root to the leafs. The distribution of the new class is restricted to a small area in feature space defined by the single leaf node reached by the training sample. It is obvious, that this leads to an over-fitting situation. Therefore we propose to use MAP estimation of leaf probabilities in terms of (2), which allows to incorporate prior knowledge as a well defined distribution. Our MAP estimation technique, as well as the underlying model distributions, are presented in section 4.

3.3 Leaf probabilities as a parameter

The event of an example reaching node n is defined by the probability mass of a part of the feature space, described by the path from the root to n. For this reason, we denote this event by Ω^n analogous to Ω_i representing the part of the feature space related to class i. If the sum of training example weights reaching each node $c(\Omega_i|\Omega^n)$ (non-normalized posterior distribution, sum of weights) is stored, node posterior distributions $p(\Omega_i|\Omega^n)$ are easily converted to node probabilities $p(\Omega^n|\Omega_i)$ using the following recursive formula (p parent node of n):

$$\begin{aligned} p(\Omega^n \mid \Omega_i) &= p(\Omega^p \mid \Omega_i)\; p(\Omega^n \mid \Omega^p, \Omega_i) \\ &= p(\Omega^p \mid \Omega_i)\; \left(\frac{c(\Omega_i \mid \Omega^n)}{c(\Omega_i \mid \Omega^p)} \right) . \end{aligned} \quad (3)$$

The trivial case is the probability of the root node r: $p(\Omega^r \mid \Omega_i) = 1$. Our parameter vector $\boldsymbol{\theta}$ for MAP estimation consists of the node probability of each leaf l:

$$\theta_l = p(\Omega^l \mid \Omega_\gamma) . \quad (4)$$

Additionally we use $\boldsymbol{\theta}^i$ to denote the vector of leaf probabilities corresponding to an arbitrary class i. Due to the fact that leafs of a decision tree induce a partitioning of the feature space in disjoint subsets Ω^l, each instance of the parameter vector is a discrete multinomial distribution. For this reason any suitable distribution of discrete distributions can be used to model a prior and perform MAP estimation. In section 4 we present our model of the prior as well as the resulting MAP estimation of discrete distributions. At this point we assume that the leaf probabilities of the new class are well estimated. Inner node probabilities can be calculated additionally by simple recursive summation within the tree. The last step is to recompute all posterior distributions, which can be achieved by the usual application of Bayes' law:

$$\begin{aligned} p(\Omega_i \mid \Omega^n) &= \frac{p(\Omega^n \mid \Omega_i)\; p(\Omega_i)}{p(\Omega^n)} \\ &= \frac{p(\Omega^n \mid \Omega_i)\; p(\Omega_i)}{\sum_{i'} p(\Omega^n \mid \Omega_{i'})\; p(\Omega_{i'})} . \end{aligned} \quad (5)$$

3.4 Additional Discriminative Levels

All machine learning approaches using the inter-class paradigm within a single classification task have to cope a common issue: Transferring knowledge from support classes can lead to confusion with the new class. For example using prior information from camel images to support a dromedary within an animal classification task enables to transfer shared features like fur color or head appearance. However, the classifier has to use additional features like shape information to discriminate between the two object categories.

To solve this problem we propose to build additional discriminative levels of the decision tree after

MAP estimation of the leaf posterior distributions. Starting from a leaf node with non-zero posterior probability of the new class, we use the method of [10] to further extend the tree. Training data in this case consists of all samples of the new class and samples of all other classes which reached the leaf. All of the samples are weighted by the corresponding unnormalized posterior distribution. This procedure allows to find new discriminative features especially between the new class and all support classes.

4 Constrained Gaussian Prior

The selection of a suitable parametric model for the prior distribution is a difficult task. The model parameter, which is estimated, is a discrete distribution itself, thus the prior distribution is additionally constrained to a $(n-1)$-simplex. A common choice is a conjugate prior, as a Dirichlet distribution for multinomial distributions. This family of parametric distributions has typically as many hyper-parameters as parameters of the underlying problem. Therefore one needs a huge set of samples from support classes to estimate optimal Dirichlet parameters.

For this reason we propose to use a constrained Gaussian distribution (CGD), which is a much simpler family of parametric distributions. With $\boldsymbol{\theta} \geq 0$ we define the density as:

$$p(\boldsymbol{\theta}|T^{\mathcal{S}}) \propto \mathcal{N}(\boldsymbol{\theta}|\,\boldsymbol{\mu}^{\mathcal{S}}, \sigma^2\mathbf{I})\,\delta\left(1-\sum_l \theta_l\right). \quad (6)$$

We multiply with the δ-term ($\delta(0) = 1, \forall x \neq 0 : \delta(x) = 0$) to ensure, that the support of the density function is the simplex with all feasible discrete distributions. The use of a $\sigma^2\mathbf{I}$ as a covariance matrix is an additional assumption which is useful to derive an efficient MAP estimation algorithm (section 4.1).

Figure 3 presents a graphical comparison between a Dirichlet distribution and our CGD for some parameter values. Note that our model can approximate a Dirichlet for the whole variety of parameters. The main difference is the elliptic shape of iso-lines in contrast to a more triangular shape, which is of course more appropriate for the simplex.

This simple model allows to estimate hyper-parameters $\boldsymbol{\mu}^{\mathcal{S}}, \sigma$ in an usual way and efficiently perform MAP estimation. The mean vector $\boldsymbol{\mu}^{\mathcal{S}}$ can be estimated analogous to a non-constrained Gaussian, because of the simplex being a convex set.

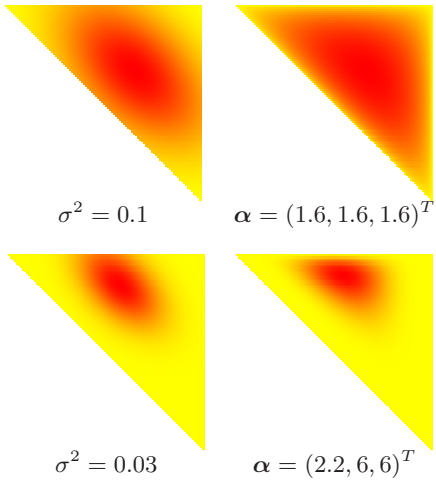

Figure 3: Comparison between a constrained Gaussian (left column) and a Dirichlet distribution (right column). Color represents the value of the density at the position on the simplex. Mean of the CGD is set to the mean of the corresponding Dirichlet distribution

Within our application on decision trees, $\boldsymbol{\mu}^{\mathcal{S}}$ is estimated using leaf probabilities of support classes:

$$\boldsymbol{\mu}^{\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \boldsymbol{\theta}^i\,. \quad (7)$$

As usual, our model of the unknown distribution by a gaussian parametric family is mostly related to computational practical issues rather than theoretical results. Applied to our leaf probabilities with regularization using support classes, this simple model can of course lead in some cases to a wrong estimation. For example support classes could share a common feature which is not related to the new class. In spite of these possible failure cases, we will show that due to the high dimensionality of $\boldsymbol{\theta}$ and careful averaging using bagging our simple gaussian model is sufficient to increase the performance of different classification tasks.

4.1 MAP Estimation using a CGP

MAP estimation using complex parametric distribution often needs nonlinear optimization techniques. In contrast to these approaches we show that by using our constrained Gaussian as a prior of a multinomial distribution, it is possible to derive a closed-form solution of the global optimum depending on a single Lagrange multiplier.

We start by writing the objective function of the MAP estimation as a Lagrange function of our simplex constraint and the posterior:

$$L(\boldsymbol{\theta}, \lambda) = \log\left(p(T^\gamma|\boldsymbol{\theta})\; p(\boldsymbol{\theta}|T^\mathcal{S})\right) + \lambda\left(\sum_l \theta_l - 1\right) . \tag{8}$$

The likelihood has the simple form of a multinomial and depends on a discrete histogram $\mathbf{c} = (c_l)_{l=1}^m$ representing the number of samples of each component:

$$p(T^\gamma|\boldsymbol{\theta}) \propto \prod_l (\theta_l)^{c_l} . \tag{9}$$

Within our application to leaf probabilities of decision trees, the absolute number of examples reaching a node is used: $c_l = c(\Omega_\gamma|\Omega^l)$. Hence, the overall objective function can be written as:

$$\sum_l \left(c_l \log(\theta_l) - \frac{1}{2\sigma^2}(\theta_l - \mu_l)^2 + \lambda\theta_l\right) - \lambda .$$

The normalization factor of our prior and the likelihood can be neglected, because they are single independent additive constants in L. Setting the gradient $\left(\frac{\partial L}{\partial \theta_l}\right)(\boldsymbol{\theta}, \lambda)$ to zero leads to m independent equations:

$$0 = \frac{c_l}{\theta_l} - \frac{1}{2\sigma^2} \cdot 2 \cdot (\theta_l - \mu_l) + \lambda . \tag{10}$$

Note that we get a non-informative prior which reduces MAP to ML estimation with $\sigma^2 \to \infty$. It is easy to proof, that under "non-degenerate conditions" all entries of the optimal vector $\boldsymbol{\theta}^{\text{MAP}}$ are positive. Therefore we can assume $\theta_l > 0$ for each l and it is possible to obtain a simple quadratic equation in θ_l:

$$0 = \theta_l^2 + \theta_l\,(-\mu_l - \lambda\sigma^2) - \sigma^2 c_l . \tag{11}$$

Therefore the optimization problem only has a single positive solution depending on λ:

$$\theta_l = \frac{\mu_l + \lambda\sigma^2}{2} + \sqrt{\left(\frac{\mu_l + \lambda\sigma^2}{2}\right)^2 + \sigma^2 c_l} . \tag{12}$$

Estimating the Lagrange multiplier is done with a simple fixed point iteration, derived from our simplex constraint:

$$\begin{aligned} 1 &\stackrel{!}{=} \sum_l \theta_l \\ &= \sum_l \frac{\mu_l + \lambda\sigma^2}{2} + \sum_l \sqrt{g_l(\lambda)} \\ &= \frac{1}{2} + \frac{1}{2} m\lambda\sigma^2 + \sum_l \sqrt{g_l(\lambda)} \end{aligned} \tag{13}$$

where

$$g_l(\lambda) = \left(\frac{\mu_l + \lambda\sigma^2}{2}\right)^2 + \sigma^2 c_l \tag{14}$$

is an abbreviation of the term under the square root. This finally leads to the following recursion formula:

$$\lambda^{i+1} = \frac{1 - 2\sum_l \sqrt{g_l(\lambda^i)}}{m\sigma^2} . \tag{15}$$

So far, we cannot prove convergence theoretically, but in our application we find a suitable solution within a few iterations. Better techniques to solve equation (13) beyond a simple fixed point iteration may lead to better numerical stability but are not investigated. Given the Lagrange multiplier λ we can easily estimate the whole vector $\boldsymbol{\theta}$ using equation (12).

5 Experiments

To show its applicability, we experimentally evaluated our approach. For a comparative analysis the use of publicly available datasets is an important aspect. On that account the database of handwritten Latin letters [7] and the demanding Mammals database [9] are used. The diversity of the two databases allows to assess the general performance

gain of our method independent of the classification task.

Evaluation criterions are unbiased average recognition rates of the whole classification task and single recognition rates of the new class. We perform Monte Carlo analysis by selecting randomly f training examples of the new class, which leaves the rest for testing. To estimate recognition rates for a fixed value of f we average the results of multiple runs. This also averages the influence of the highly randomized manner of our base classifier.

Our experimental evaluation aims at analyzing the gain of our transfer learning approach compared to a standard ERT classifier. We do not focus on the development of new feature types which are suitable for a special recognition task. For this reason our choice of features is quite standard and not optimized (section 5.1 and 5.2).

The variance σ^2 of the CGP is an important parameter of our method. It controls the influence of the prior and therefore indirectly our assumption about how much the new class is related to support classes. An optimal value for σ^2 could be obtained by cross validation, but we fixed the value to 10^{-5} in all experiments.

Another issue is the selection of support classes from all available classes. Our main assumption in equation (2), suggests that those categories have to share common features, shape, appearance or context (section 5.3). Automatically estimating class similarity would be optimal to provide support class subsets. However this leads to a vicious circle, because one has to estimate models in advance. Therefore we select support classes manually.

We build an ensemble of 10 decision trees without limits on depth or example counts within each node. During training we test a maximum of 500 random features with 15 thresholds drawn from a uniform distribution at each node.

5.1 Latin Letters

The database of [7] is a collection of images containing handwritten Latin letters resulting in 26 object categories. For each object class 59-60 images are provided.

Features Images of this database are binary, so a very simple feature extraction method is used. We divide the whole image into an equally spaced $w_x \times w_y$ grid. In each cell of the grid, the ratio of black pixels and all pixels within the cell is used as a single feature. This leads to a feature vector with $w_x w_y$ dimensions. In all experiments we used values of $w_x = 8$ and $w_y = 12$.

amur tiger	red panda	alpaca
hippopotamus	baboon	tapir
sea lion	black lemur	*hyena*
asian elephant	caracal	opossum
beluga whale	siberian tiger	ocelot
hartebeest	cape buffalo	moose
african elephant	ferret	llama
howler monkey	bighorn sheep	marmot
bengal tiger	*african lion*	lemur
white rhinoceros		

Table 1: Categories used for experiments with the Mammals database[9]: bold font = class with very few training examples, italic font = support classes

5.2 Mammals Database

The Mammals database of Fink and Ullmann [9] consists of over 45000 images categorized into 409 animal types. Categorization of this database is very difficult, because of large intraclass variability, included art drawings and mixed head and body images. In all our experiments we used a subset of all object categories in the database (Table 1).

The database was specially collected to push progress in the area of object recognition using the interclass transfer paradigm. Many object categories share common features due to evolutionary relationships. We selected support categories with rather generic properties shared with the new class "sea lion". Therefore our results may provide a lower bound on performance gain, that is achieved with very similar support classes. It is important to note, that we did not manually align and flip images or use bounding box information.

Features Due to large intraclass variations global features are not suitable for this classification task. Therefore we build upon the work of [16]. Standard SIFT descriptors $\mathbf{d}_k^{\mathcal{I}}$ are evaluated at several random locations within the image $\mathcal{I}$ (e.g. $g = 1000$ positions). All descriptors are used to train the classifier. Classification of an image $\mathcal{I}$ is done by simple averaging all descriptor posteriori probabilities:

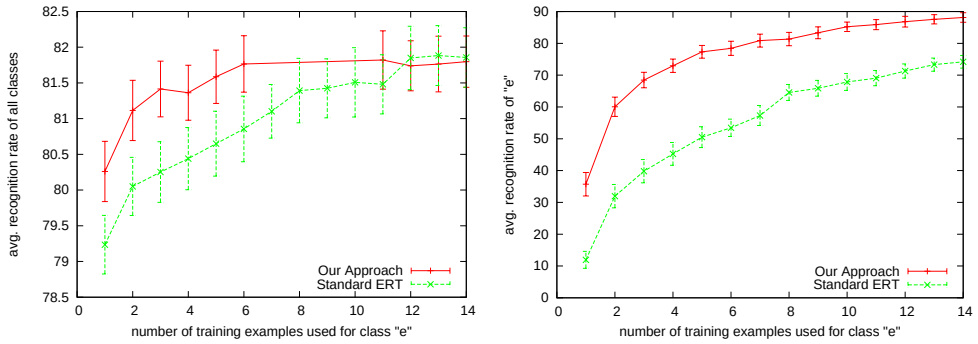

Figure 4: Results for the Latin Letters Database of Fink et al. [7]: recognition rates of our classifier and the standard approach applied to handwritten Latin letters with few training examples for the letter "e" (support classes: "a,c,d,b", 30 training examples used for all other classes). Note that due to the high randomization, error bars display 0.25σ ranges.

$$p(\Omega_i \mid \mathcal{I}) = \frac{1}{g} \sum_{k=1}^{g} p(\Omega_i \mid \mathbf{d}_k^{\mathcal{I}}) \; . \qquad (16)$$

Features are based on gray-values and no incorporation of color information is done. We also tested a codebook approach [18] with codebook size 1000 obtained by online k-Means. This method only reached about 17% average recognition rate compared to 22% achieved by the method of [16].

5.3 Transferring context information

The success of orderless Bag-of-Features approaches (BoF) as a standard method for object classification can be traced back to two aspects: First of all geometric models are difficult to learn and current methods are not yet robust enough to handle large intraclass variation. Thus, order-less methods provide more robust classifiers. Another advantage of BoF methods, which seems to be the most important one, is the indirect use of context information. Background and foreground features are used equally in all steps of the training and classification process. For this reason, a huge amount of background features influences a BoF method.

Our method naturally transfers common features, which can be object-specific or contextual, to a new class. For example it could transfer the knowledge of typical desert-like background from a camel class to a dromedary class.

5.4 Evaluation

Results of the comparison between our approach and the standard ERT classifier [10] are presented in Figure 4 and 5. Both average recognition rates of the whole classification task and recognition rates of the new class are plotted as a function of training examples. Our combined generative/discriminative transfer learning approach outperforms the standard ERT classifier when **using very few training examples**. This shows, that improving recognition of the new class does not reduce the classification performance of other classes on average.

By training with a single example of the Latin letter "e", a recognition rate of 35.71% was reached on average, compared to 11.93% by the standard ERT classifier. Using two training examples the gap even increases from 60% of our approach to 31% of the standard one.

An important aspect of our method can be seen in both of the experiments. After a certain number of training images used for the new class MAP estimation does not improve average recognition rates of the whole classification task but increases classification accuracy of the new class. Thus, the performance on other classes decreases due to confusion with the new class. This can be traced back to the fact, that the prior has to much weight in spite of a very informative likelihood.

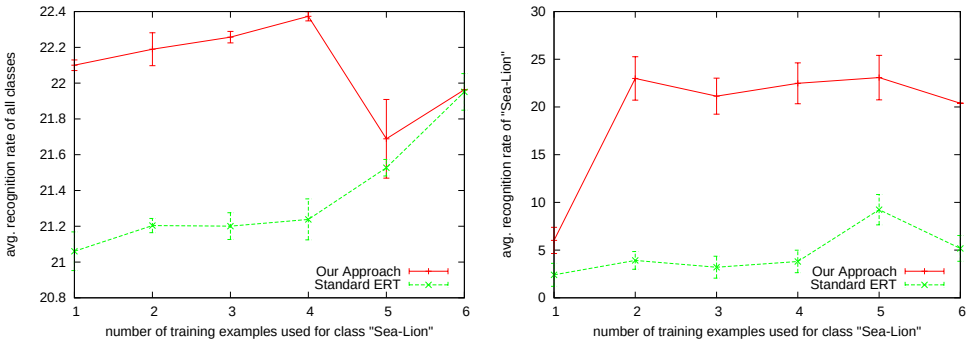

Figure 5: Results for the very difficult Mammals Database [9] with 28 categories: recognition rates of our classifier and the standard approach applied to images of mammals with few training examples for the class "sea lion" (support classes: african lion, beluga whale, hyena, hippopotamus, 50 training examples used for all other classes). Note that due to the high randomization, error bars display 0.25σ ranges.

6 Conclusion

We argue that learning with few examples can be solved by incorporating prior knowledge of related classes (interclass transfer paradigm). For this reason, a combined discriminative/generative classifier was presented that uses interclass relationships to support object classes with few training examples.

As a first step, a discriminative randomized decision tree classifier is learned from all classes. The key concept of our approach is a subsequent MAP estimation of leaf probabilities within each tree for one class with weak training representation. Bayesian formulation allows to infer knowledge from related classes as prior distribution, which needs a suitable parametric model. Therefore we introduced a new family of prior distributions for a multinomial distribution, which is a Gaussian constrained to a simplex. Resulting MAP estimation leads to easily solvable equations without the need of complex nonlinear optimization techniques.

Experiments performed on two public available datasets (Latin letters of [7], Mammals database [9]) show how our method applied to a classification task significantly boosts the recognition rate of one class with very few examples. In case of the Latin letters database we were able to improve recognition using a single training example of the class "e" about 23%. Experimental evaluation using the Mammals database show that our method is able to reach a gain of up to 19% using two training examples within a very difficult classification task.

7 Further Work

Our compact model of a prior for multinomial distributions can be applied beyond our hybrid classifier. Thus, we are interested in which situations a CGP is more efficient and suitable than a common Dirichlet distribution.

One issue of our algorithm is of course a manually selected variance of the prior, which controls the influence of related classes. Therefore another question of interest is whether it is possible to estimate σ automatically with all information about the classification task one can extract from training images.

Acknowledgments We would like to thank ROBOT Visual Systems GmbH for financial support and all of our colleagues for valuable comments.

References

[1] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 17–24, New York, NY, USA, 2007. ACM Press.

[2] Leo Breiman. Random forests. *Machine Learning*, V45(1):5–32, October 2001.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[4] Li Fei-Fei. Knowledge transfer in learning to recognize visual objects classes. In *Proceedings of the International Conference on Development and Learning (ICDL)*, 2006.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594, 2006.

[6] Andras Ferencz, Erik G. Learned-Miller, and Jitendra Malik. Building a classification cascade for visual identification from one example. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 286–293, Washington, DC, USA, 2005. IEEE Computer Society.

[7] Michael Fink. Object classification from a single example utilizing class relevance pseudometrics. In *Advances in Neural Information Processing Systems*, volume 17, pages 449–456. The MIT Press, 2004.

[8] Michael Fink and Pietro Perona. Mutual boosting for contextual inference. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schoelkopf, editors, *Advances in Neural Information Processing Systems*. MIT Press, 2003.

[9] Michael Fink and Shimon Ullman. From aardvark to zorro: A benchmark for mammal image classification. *Int. J. Comput. Vision*, 77(1-3):143–156, 2008.

[10] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, 2006.

[11] D. Hoiem, C. Rother, and J. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07.*, pages 1–8, 2007.

[12] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *International Conference of Computer Vision (ICCV)*, volume 1, pages 654 – 661. IEEE, October 2005.

[13] Alex D. Holub, Max Welling, and Pietro Perona. Hybrid generative-discriminative visual categorization. *Int. J. Comput. Vision*, 77(1-3):239–258, 2008.

[14] Susan S. Jones and Linda B. Smith. The place of perception in children's concepts. *Cognitive Development*, 8:113–139, April-June 1993.

[15] Vincent Lepetit, Pascal Lagger, and Pascal Fua. Randomized trees for real-time keypoint recognition. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 775–781, Washington, DC, USA, 2005. IEEE Computer Society.

[16] Raphaël Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel. Random subwindows for robust image classification. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 34–40. IEEE, June 2005.

[17] Erik G. Miller, Nicholas E. Matsakis, and Paul A. Viola. Learning from one example through shared densities on transforms. In *CVPR*, volume 1, pages 464–471, Los Alamitos, CA, USA, 2000. IEEE Computer Society.

[18] Eric Nowak, Frederic Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *European Conference on Computer Vision*, volume 3954 of *Lecture Notes in Computer Science*, pages 490–503. Springer, 2006.

[19] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pages 1–8, Minneapolis, Minnesota, USA, June 2007. IEEE Computer Society.

[20] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR 2008*, 2008. accepted for publication.

[21] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):854–869, 2007.