Learning with Few Examples for Binary and Multiclass Classification Using Regularization of Randomized Trees

Erik Rodner*, Joachim Denzler

Chair for Computer Vision, Friedrich Schiller University of Jena, Germany

Abstract

The human visual system is often able to learn to recognize difficult object categories from only a single view, whereas automatic object recognition with few training examples is still a challenging task. This is mainly due to the human ability to transfer knowledge from related classes. Therefore, an extension to Randomized Decision Trees is introduced for learning with very few examples by exploiting interclass relationships. The approach consists of a maximum a posteriori estimation of classifier parameters using a prior distribution learned from similar object categories. Experiments on binary and multiclass classification tasks show significant performance gains.

Key words: object categorization, randomized trees, few examples, interclass transfer, transfer learning

1. Introduction

During the last few decades, research in machine learning and computer vision has led to many new object representations and improved algorithms for numerical classification. Despite the success of this development, there is still an unanswered question: how does one learn object models from few training examples?

On the one hand, this question is motivated from industrial demand. In many applications, gathering hundreds or thousands of training images is either expensive or nearly impossible (Platzer et al., 2008). Building robust classification systems in those settings therefore requires complex specialized methods, that indirectly incorporate human prior knowledge about the task.

On the other hand, progress on learning with few examples is an important challenge and an essential step towards closing the gap between human and computer vision abilities. The human visual recognition system is often easily able to learn a new object category, such as a new animal class, from just a single view.

At first glance, this observation seems to contradict to the classical theory. The parameters of object models often exceed the available number of training examples in multiple dimensions. From a mathematical point of view, this results in an ill-posed optimization problem, especially in cases with only a few training examples.

Therefore the only possibility to solve this problem is to regularize the optimization using prior knowledge. In previous algorithms this prior knowledge was often derived from abstract assumptions or was manually tuned during the development. However psychological studies (Jones and Smith, 1993) suggest that a key component of the human ability to recognize a class from a limited number of examples is the concept of *interclass transfer*. This paradigm is also known as *knowledge transfer*, *learning to learn* or *transfer learning*. It states that prior knowledge from previously learned object categories is the most important additional information source when learning object models from weak representations (Fei-Fei, 2006). To give an illustrative example of this idea, consider the recognition of a new animal class such as an okapi. With the aid of our prior knowledge from related animal classes (giraffe, zebra, antelope, etc.), we are able to generalize quickly from a single view.

In this work, a concept is presented how prior knowledge of related classes (often also called support classes) can be used to increase the generalization ability of a discriminative classifier. The underlying idea is a maximum a posteriori (MAP) estimation of parameters using a prior distribution estimated from similar

*Corresponding author

Email addresses: rodner@informatik.uni-jena.de (Erik Rodner), denzler@informatik.uni-jena.de (Joachim Denzler)

object categories. Furthermore, the application of this idea to Randomized Decision Trees, as introduced by Geurts et al. (2006), is demonstrated. The paper is based on our previous work in Rodner and Denzler (2008) that concentrates on multiclass classification. Studies are extended by showing the applicability of the approach to binary classification. An additional experiment also emphasizes that the information transferred is not generic prior knowledge unrelated to interclass relationships.

The remainder of the paper is organized as follows. After previous work in the field of learning with weak representations is briefly reviewed, it is shown that Bayesian estimation using a prior distribution is a well founded possibility to transfer knowledge from related classes (Bayesian Interclass Transfer). This is followed by a detailed description of an extension to Randomized Decision Trees in Section 4, which can be regarded as an application of Bayesian Interclass Transfer. Experiments in binary and multiclass classification settings using publicly available image databases demonstrate the benefits of the proposed algorithm in Sections 6 to 9. A summary of our findings and a discussion about further research steps conclude the paper.

2. Related Work

Previous work on interclass transfer varies significantly in the type of information transferred from related classes. An intuitive assumption is that similar classes share common intraclass geometric transformations. The *Congealing* approach of Miller et al. (2000) therefore tries to estimate those transformations and use them to increase the amount of training data of a new class. For example, a single training image of a letter in a text recognition setting can be transformed using typical rotations estimated from other letters.

Another idea is to assume shared structures in feature space and estimate a metric or transformation from support classes (Fink, 2004; Quattoni et al., 2007). Torralba et al. (2007) used a discriminative boosting technique that exploits shared class boundaries within feature space. In contrast, Fei-Fei et al. (2006) developed a generative framework with MAP estimation of model parameters using a prior distribution estimated from support classes. A similar idea in the context of shape based image categorization is presented in Stark et al. (2009). In general the concept of shared priors for a set of related classification problems can be used to extend several classification techniques to multi-task approaches, such as generalized linear models (Lee et al., 2007) or Gaussian processes (Bonilla et al., 2008).

Our work on regularized decision trees using transfer learning is related to the approach of Lee and Giraud-Carrier (2007). The key idea of their method is the reusability of a decision tree structure from a related binary classification task. In contrast, this paper introduces a technique that also reuses estimated class probabilities in leaf nodes and performs a re-estimation based on a Bayesian framework.

3. Bayesian Interclass Transfer

The interclass transfer paradigm leads quickly to two important questions: What type of information can be transferred, and how can this be done using machine learning techniques? The first question is answered in Section 4.2. Here we concentrate on the description of how prior knowledge can be incorporated.

Let a set $\mathcal{S}$ of support classes and a class γ with few training examples be given. In the remainder of this paper, class γ is called the new class. The overall goal of Bayesian Interclass Transfer is to estimate an object model $\boldsymbol{\theta}(\gamma)$ (parameters of a distribution, parameters of a classifier, etc.) with the help of prior knowledge from related object models $\boldsymbol{\theta}(i)$ where $i \in \mathcal{S}$. Using the Bayesian principle, this can be formulated as the following maximum a posteriori estimation

$$\boldsymbol{\theta}^{\mathrm{MAP}}(\gamma) = \arg\max_{\boldsymbol{\theta}} p(T^{\gamma} \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta} \mid T^{\mathcal{S}}) \;, \qquad (1)$$

where T^{γ} denotes the training data of the new class and $T^{\mathcal{S}}$ denotes the training data of all support classes. The fundamental assumption is that it is possible to estimate a suitable prior distribution and use it to regularize the parameter estimation of a related class.

The application of the principle of Bayesian Interclass Transfer (or Generative Transfer Learning) was limited to generative approaches (Fei-Fei et al., 2006). As we show in this paper it is also possible to enhance a discriminative classifier. The key idea is the re-estimation of parameters of a discriminative classifier by MAP estimation.

For this reason we propose to estimate the parameters $\boldsymbol{\theta}(i)$ $(i \neq \gamma)$ using a state-of-the-art discriminative approach and only recompute the parameters of the new class $\boldsymbol{\theta}(\gamma)$ with further regularization. Figure 1 gives an overview of this concept.

4. Regularized Randomized Trees

This section describes how to apply the previous idea of Bayesian Interclass Transfer to decision tree classifiers. Although the approach can be easily applied to

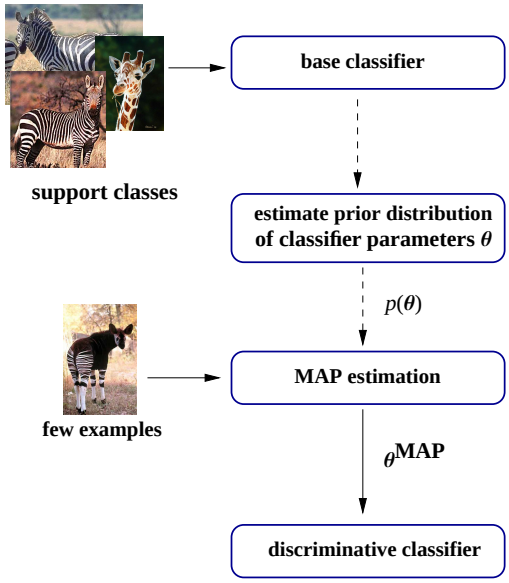

Figure 1: Overview of our approach using Bayesian Interclass Transfer for parameter estimation within a discriminative classification approach.

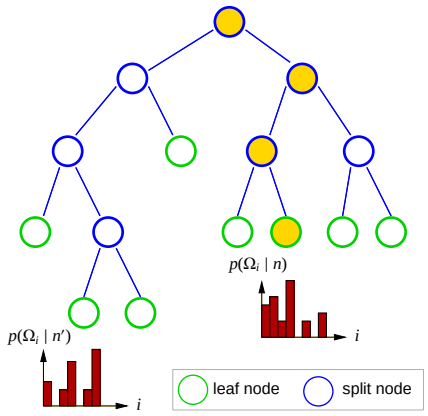

Figure 2: General principle and terms of decision trees. Diagrams illustrate the posterior distribution within each leaf node. Traversal of the tree (nodes filled with grey/yellow color) is done using features stored within each split node.

arbitrary decision tree approaches, the Randomized Decision Forest (RDF) approach is used, because of its superior generalization performance and its widely use in different applications (Marée et al., 2005; Shotton et al., 2008). In this section, we review RDF before providing a step-wise description of our method.

4.1. Randomized Decision Trees

Decision tree classifiers are commonly binary trees with two types of nodes. Each inner node represents a weak classifier (one-dimensional feature and threshold) which defines a hyperplane in feature space and thus determines the traversal of a new example within the tree. The traversal of the tree ends in a leaf node n.

We use n or n^l with $l = 1 \ldots m$ to denote the event of an example reaching a single leaf node of a decision tree. This event also corresponds to the infinite set of all such examples (feature vectors). The total number of all leaf nodes in a single decision tree is denoted by m. Each leaf node is associated with a posterior distribution $p(\Omega_i \mid n)$, which is an estimate of the probability of class i given that this specific leaf is reached. We denote by Ω_i the event of an example belonging to the class i. These general principles and terms are illustrated in Figure 2.

Standard decision tree approaches suffer from two serious problems: long training time and over-fitting. The RDF approach solves both issues by random sampling. Instead of evaluating every feature and threshold, the training time is reduced by an approximate search for the most informative weak classifier in each node. The selection is made 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, the generalization performance can be improved by learning an ensemble of M decision trees (often called a forest) using a random subset of the training data. From the final leaf nodes of the forest $\mathbf{n} = (n_1, \ldots, n_M)$, the overall posterior can be obtained by voting with equal weights:

$$p(\Omega_i \mid \mathbf{n}) = \frac{1}{M} \sum_{s=1}^{M} p(\Omega_i \mid n_s) \ . \qquad (2)$$

This special case of Bagging (Breiman, 2001) reduces the over-fitting effects without the need for additional tree pruning.

4.2. Transfer Learning Using RDF

The transfer learning idea can be applied to each tree of the forest individually; therefore, the details of our method are explained using only a single decision tree. Two different types of information are transferred: a discriminative tree structure and a prior distribution on leaf probabilities.

4.2.1. Recycling of Decision Trees

The selection of discriminative features in high dimensional spaces using few examples is a highly ill-posed problem. Therefore, we construct a discriminative tree structure using all the available training data of

all classes. This concept has also been used in Hoiem et al. (2007) and Lepetit et al. (2005), to recycle features and to reduce computation time. The assumption of shared discriminative features (or weak learners) is closely related to the use of shared features in the work of Torralba et al. (2007).

4.2.2. *Re-estimation of leaf probabilities*

Although decision tree approaches can be considered as discriminative, they are closely related to individual density estimation. The tree structure is a partitioning of the whole feature space into several cells n^l represented by leaf nodes. This corresponds to an approximation of a class distribution using a piecewise constant density or discrete probability distribution. The leaf probabilities $\theta_l^i = p(n^l \mid \Omega_i)$ are the maximum likelihood (ML) estimates of a multinomial distribution estimating the density of each cell:

$$\theta_l^{\mathrm{ML}}(i) = \frac{|n^l \cap T^i|}{|T^i|} \quad . \tag{3}$$

Note that $|n^l \cap T^i|$ is the number of examples of class i reaching a node n_l during the training step. It should be noted that with a careful implementation of decision trees, which store those unnormalized values instead of the posterior probability, a complicated recursive computation of leaf probabilities as presented in Rodner and Denzler (2008) is not necessary.

It is obvious that with only a few training examples $\mathbf{x} \in T^\gamma$, the vector $\boldsymbol{\theta}^{\mathrm{ML}}(\gamma)$ is sparse and is unable to provide a good approximation of the underlying distribution. The overall goal of our approach is to re-estimate $\boldsymbol{\theta}(\gamma)$ by using maximum a posteriori estimation, which leads to a smoother solution $\boldsymbol{\theta}^{\mathrm{MAP}}(\gamma)$. Since the leaves of a decision tree induce a partitioning in disjoint subsets n^l, each instance of the parameter vector $\boldsymbol{\theta}$ is a discrete multinomial distribution. For this reason any suitable distribution of discrete distributions can be used to model the prior distribution.

4.3. *Constrained Gaussian Prior*

We propose to use a constrained Gaussian distribution (CGD), which is a simple family of parametric distributions and can serve as an alternative to a standard Dirichlet distribution. For all $l : \theta_l \geq 0$ the density is defined as

$$p(\boldsymbol{\theta}|T^{\mathcal{S}}) \propto \mathcal{N}(\boldsymbol{\theta}|\,\boldsymbol{\mu}^{\mathcal{S}}, \sigma^2\mathbf{I})\,\delta\left(1 - \sum_l \theta_l\right) \quad . \tag{4}$$

The factor of δ ($\delta(0) = 1, \forall x \neq 0 : \delta(x) = 0$) is essential to ensure that the support of the density function is the simplex of all feasible discrete distributions. The use of $\sigma^2\mathbf{I}$ as a covariance matrix is an additional assumption that will be useful in deriving an efficient MAP estimation algorithm (Section 4.4).

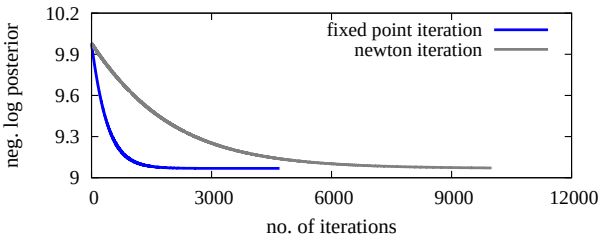

Figure 3: Comparison between the convergence of the newton method and a simple fixed point iteration.

This simple model allows us to estimate hyperparameters $\boldsymbol{\mu}^{\mathcal{S}}$ and σ in an usual way. Because the simplex is a convex set, the mean vector $\boldsymbol{\mu}^{\mathcal{S}}$ can be estimated analogously to a non-constrained Gaussian. In our application on decision trees, $\boldsymbol{\mu}^{\mathcal{S}}$ is estimated using the leaf probabilities of the support classes:

$$\boldsymbol{\mu}^{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \boldsymbol{\theta}(i) \quad . \tag{5}$$

Our choice to model the unknown distribution by a Gaussian parametric family is mostly due to practical computational considerations rather than theoretical results. Of course, one could argue, that using a symmetric Dirichlet prior leads to the same set of parameters as a CGD and is additionally a conjugate prior. In our application for Regularized trees, we expect a symmetric Dirichlet prior to yield similar results. But in our opinion the use of a constrained Gaussian prior is scientifically interesting and we will show in the following that even without a conjugate prior, one can derive a simple inference method using an easy to solve one-dimensional optimization problem. An investigation and analysis of other parametric distributions and more sophisticated priors would be an interesting topic for future research.

4.4. *MAP Estimation using a CGP*

The process of MAP estimation using complex parametric distribution often requires nonlinear optimization techniques. In contrast to these approaches we briefly 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:

$$L(\boldsymbol{\theta}, \lambda) = \log\left(p(T^\gamma|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|T^\mathcal{S})\right) + \lambda\left(\sum_l \theta_l - 1\right) . \quad (6)$$

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

$$p(T^\gamma|\boldsymbol{\theta}) \propto \prod_l (\theta_l)^{c_l} \quad . \quad (7)$$

In our application to leaf probabilities of decision trees, the absolute number of examples reaching a node $c_l = |n^l \cap T^\gamma|$ is used, where m is the number of all leaves. With the CGD prior in equation 4 we obtain the overall objective function

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

This objective function is convex and therefore has a unique solution. Setting the gradient $\left(\frac{\partial L}{\partial \theta_l}\right)(\boldsymbol{\theta}, \lambda)$ to zero leads to the m independent equations

$$0 = \frac{c_l}{\theta_l} - \frac{1}{2\sigma^2} \cdot 2 \cdot (\theta_l - \mu_l) + \lambda \quad . \quad (8)$$

Note that we get a non-informative prior, which reduces MAP to ML estimation as $\sigma^2 \to \infty$. With positive discrete probabilities ($\theta_l > 0$), it is possible to obtain a simple quadratic equation in θ_l:

$$0 = \theta_l^2 + \theta_l\ (-\mu_l - \lambda\sigma^2) - \sigma^2 c_l \quad . \quad (9)$$

A stationary point with $\theta_l = 0$ is only possible with $c_l = 0$ or $\sigma^2 \to 0$, which is also reflected by the above equation. Therefore the optimization problem has only a single non-negative solution depending on λ:

$$\theta_l = \frac{\mu_l + \lambda\sigma^2}{2} + \sqrt{\left(\frac{\mu_l + \lambda\sigma^2}{2}\right)^2 + \sigma^2 c_l} \quad . \quad (10)$$

This solution depends on the Lagrange multiplier, for which an optimal value can be found using a simple fixed point iteration:

$$\lambda^{j+1} = \frac{1}{m\sigma^2}\left(1 - 2\sum_l \sqrt{\left(\frac{\mu_l + \lambda^j\sigma^2}{2}\right)^2 + \sigma^2 c_l}\right) \quad . \quad (11)$$

As an initial value, it is possible to use the optimal Lagrange multiplier in the case of no prior knowledge and maximum likelihood estimation. Figure 3 shows the convergence of our technique compared to that of a Newton iteration, which converges much slower than our simple recursion formula of Equation (11).

5. Binary and Multiclass Transfer Learning

Transfer learning for binary classification relies on a set of support tasks that try to separate a class i and a background class $\mathcal{B}$. Regularized trees can be applied straight-forwardly to this setting if a single support classification task is given. After building a random forest using training data from $\mathcal{S} = \{i\}$ and $\mathcal{B}$, we can apply the re-estimation method as explained in Section 4.4 using the mean vector $\boldsymbol{\mu}^\mathcal{S} = \boldsymbol{\theta}(i)$. Finally the class probabilities of γ are substituted for all probabilities of i, so the decision tree now tries to separate between γ and $\mathcal{B}$.

In contrast to previous work, which often concentrate on the binary case (Fei-Fei et al., 2006), Regularized Trees are even suitable for multiclass classification problems. Given the leaf probabilities θ_l^i for each class i and leaf l and prior probabilities $p(\Omega_i)$ for each class, one can easily calculate the needed posterior probabilities for each class in the multiclass problem:

$$p(\Omega_i \mid n_l) = \frac{p(n_l \mid \Omega_i)\ p(\Omega_i)}{\sum_j p(n_l \mid \Omega_j)\ p(\Omega_j)} \quad . \quad (12)$$

Reducing Confusion with Support Classes. All machine learning approaches using the interclass paradigm within a multi-classification task have to cope with 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 the class dromedary enables us to transfer shared features like fur color or head appearance. However, the we have to use additional features (e.g. shape information) to discriminate between both categories.

To solve this problem, we propose to build additional discriminative levels of the decision tree after MAP estimation of the leaf distributions. Starting from a leaf node n^l with non-zero posterior probability $p(\Omega_\gamma \mid n^l)$, the tree is further extended by the randomized training procedure described in Section 4.1. The training data in this case consists of all samples of the new class and samples of all support classes which reached the leaf n^l. All of the training examples are weighted by the values of the posterior distribution $p(\Omega_i \mid n^l)$ of the leaf n^l. This technique allows us to find new discriminative

features especially between the new class and the support classes. We observed that often only one additional level can be build using the few examples of γ.

6. Experimental Setup and Overview

The approach presented is evaluated experimentally to analyze the benefits and the limitations of all our assumptions. Three experiments are performed to provide empirical proof of the following statements:

1. Regularized Trees lead to a significant performance gain for multiclass classification with few training examples (Exp. 1, Sect. 7).
2. The performance of binary classification can be improved by our method (Exp. 2, Sect. 8).
3. Our method uses prior knowledge that relies on visual similarity, and is thus not related to generic prior knowledge (Exp. 3, Sect. 9).

For the comparative analysis, three types of public datasets with different characteristics are used: a dataset of handwritten Latin letters provided by Fink (2004), a combination of the bird and butterfly datasets used in Lazebnik et al. (2004, 2006) and a dataset for binary classification using images from the database of mammals presented in Fink and Ullman (2008).

The evaluation criteria are the unbiased average recognition rates of the whole classification task and single recognition rates of the new class. Monte Carlo analysis is performed by randomly selecting f examples of the new class for training and the remainder for testing. To estimate the recognition rates for a fixed value of f the results of multiple runs are averaged. This also averages out the influence of our randomized classifier.

The experimental evaluation aims to analyze the gain of our transfer learning approach compared to the RDF classifier Geurts et al. (2006). We do not focus on the development of new feature types that would be suitable for special recognition tasks. For this reason, our choice of features is not optimized. The variance σ^2 of the CGP is an important parameter of our method, which we fix to the value of 10^{-5} in all experiments. It controls the influence of the prior distribution and therefore, indirectly, our assumption of how much the new class is related to support classes. We decided to use a constant value for this parameter, because cross-validation is impossible with a single training example.

Furthermore we select support classes manually in all the experiments. Our main assumption in Equation (1) is that those categories have to share common features, shape or appearance. Estimating the class similarities automatically would be optimal to provide support class subsets. Regarding the selection of support classes as a model selection problem allows to use cross-validation or leave-one-out estimates (cf. Tommasi and Caputo (2009)). However, this can be rather difficult and results in ill-posed problems themselves. Hence, we leave the estimation of a set of similar classes as a task for future research.

7. Experiment 1: Multiclass Classification

This experiment shows the benefits of our method in a high-level image categorization task and a simpler letter recognition task. We explain all features used and give a detailed discussion of all results in section 7.3.

7.1. Letter Recognition

The database of Fink (2004) is a collection of images containing handwritten Latin letters resulting in 26 object categories. For each object class 60 images are provided. For classification an ensemble of 10 decision trees is used and the following classification scenario is selected: new class *e* and support classes *a*,*b*,*c*,*d*.

Features. The images in this database are binary, so a very simple feature extraction method is used. The whole image is divided into an equally spaced $w_x \times w_y$ grid. In each cell of the grid, the ratio of black pixels to 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, the values $w_x = 8$ and $w_y = 12$ are used.

7.2. Image Categorization

To demonstrate the behavior of the method on a high-level image categorization task, we combine the birds (Lazebnik et al., 2006) and the butterflies dataset (Lazebnik et al., 2004) into one single multiclass classification task. The object categories can therefore be divided into two different semantic sets. The category *black swallowtail* is used as a new class γ, and all the other butterfly categories serve as support classes $\mathcal{S}$. Thus, training data consists of a variable number of training images for γ and 26 images for each of the remaining classes. This classification task is more difficult than our letter recognition setting. For this reason an ensemble of 500 decision trees was used.

Figure 4: Example images of all datasets used for experimental evaluation: Top row: combined bird and butterfly dataset of Lazebnik et al. (2006, 2004). Middle row: latin letter dataset of Fink (2004). Bottom row: zebra and okapi images used for binary classification obtained from the mammals dataset of Fink and Ullman (2008) and google image search.

Features. A standard approach to image categorization is the bag-of-features idea. A quantization of local features is computed, which is often called a codebook, is computed at the time of training. An image can then be represented as a histogram of local features with respect to codebook entries. The method of Moosmann et al. (2006), which utilizes a random forest as a clustering mechanism, is used to construct the codebook. This codebook generation procedure shows superior results compared to standard k-Means within all experiments. It also allows us to create large codebooks (a size of 13000 used in all experiments) in a few minutes on a standard PC. A combined SIFT descriptor computed on normalized RGB channels, as described in van de Sande et al. (2010), is used as a local feature representation.

7.3. Evaluation

The results of this experiment evaluating multiclass classification performance can be found in Figures 5 and 6. The plots show the average recognition rate of the whole task (plots on the left side) and the recognition rate of the new class (plots on the right side) compared to those of the original method of RDF.

It can be seen that our method improves the recognition rate of the new class and the average recognition rate, in the range with few training examples (1 to 8 examples, marked with green color). The regularization is therefore able to transfer knowledge from support classes without violating the separation between the other classes.

After a specific number of training examples the average recognition rate decreases while the recognition rate or hit-rate of class γ (plots on the right side) still grows. This critical area is highlighted in yellow in Figures 5 and 6. The effect corresponds to over-regularization. The influence of the prior distribution is controlled by hyper-parameter σ^2 which is kept a fixed value independent of the training examples used. Therefore the MAP estimation of leaf probabilities leads to many leafs with non-zero posterior probabilities for the new class. This corresponds to a large variance of the distribution in feature space, which dominates the distribution for all other classes. The variance of the class distribition reaches a critical threshold, which leads to an overestimation of the distribution corresponding to the new class. The classifier prefers the new class, which results in a worse average recognition rate (or an increasing number of false positives) on the whole classification task. It should be noted that this phenomenon is unique to our application of transfer learning in a multi-class classification task. Other transfer learning algorithms converge to the performance of independent learning after a specific number of training examples, due to their treatment of a support and new class as independent binary classification tasks. A similar effect has been observed in the context of zero-shot learning Rohrbach et al. (2010) (cf. their Fig. 3).

8. Experiment 2: Binary Classification

For an experimental evaluation of the method on a binary classification task, images from the animal categories *zebra* and *okapi* from the mammals database of Fink and Ullman (2008) are used. In order to increase the number of test images, additional images from the category *okapi* were downloaded using Google Image Search and filtered manually to delete wrong search results. The new dataset includes a total of 231 images of okapis and 200 images of zebras. The image set of the background class $\mathcal{B}$ was generated by obtaining 300

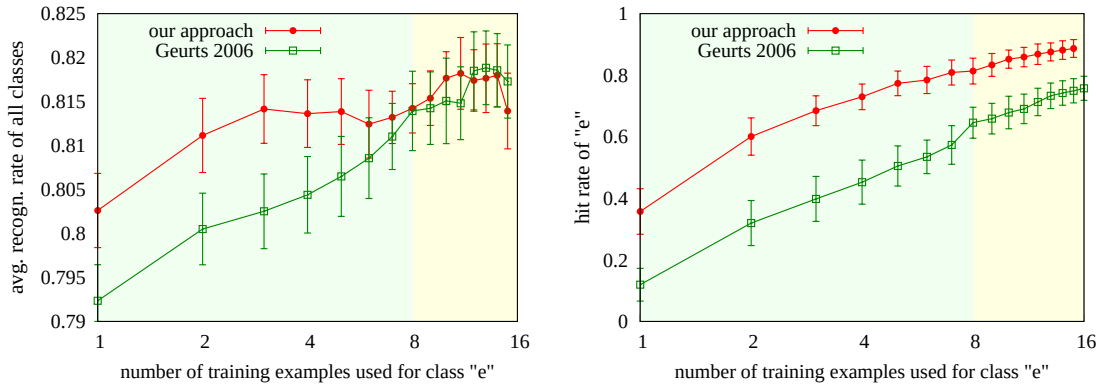

Figure 5: Comparison to Geurts et al. (2006) in a multiclass classification task using the letter recognition dataset of Fink (2004). The left plot shows the average recognition rate of the whole classification task with respect to the number of training examples (log scale) of a specific class. On the right side the single recognition rate of this class is plotted. Highlighted green area corresponds to the working range of our algorithm before over-regularization effects. False alarm rates are skipped because we concentrate on the categorization performance.

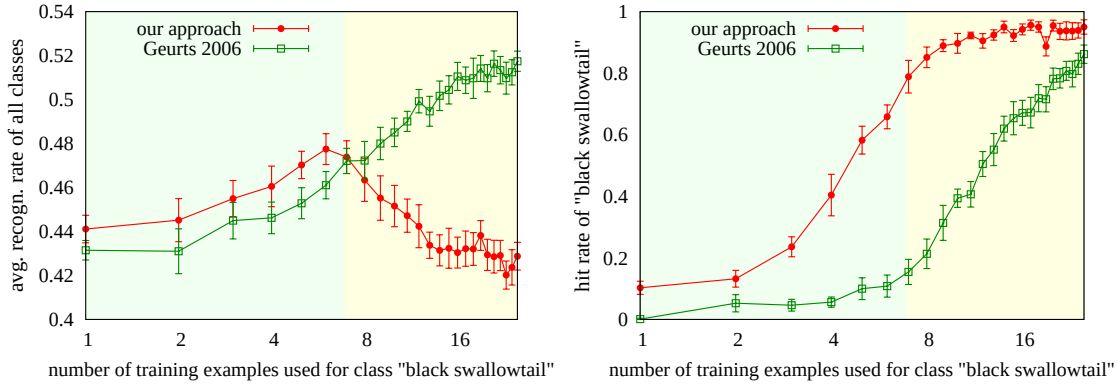

Figure 6: Comparison to Geurts et al. (2006) within a high level multiclass classification task using the bird-and-butterfly dataset as used in Lazebnik et al. (2004, 2006). Semantic of the plot is analogous to Figure 5.

random images from Google Image Search (using *the* as a search key word). Our algorithm was tested with two scenarios: using few training examples of the class okapi with the support of the class zebra and vice versa. Feature extraction was done as described in Section 7.2.

8.1. Evaluation

Figure 7 shows the results of our approach (red plot, circular dots) compared to the standard approach of Randomized Decision Forest (green plot, rectangular dots). We also tested the performance of a random forest built by using the supporting classification task without our re-estimation technique (blue plot, triangular dots).

First of all, it is apparent that our method significantly increases the classification performance compared to the standard approach in both cases. Using a random forest without re-estimation of leaf probabilities does not use training examples of the new class and is therefore independent of the number of training examples. Additionally one can see that the “okapi” task seems to be much harder, and benefits of knowledge transfer for a wider range of training examples.

9. Experiment 3: Similarity Assumption

What happens if support classes are selected that do not share common features with the new class?

As mentioned in Section 3 the concept of Bayesian Interclass Transfer is based on the main assumption that the support classes $\mathcal{S}$ are somehow similar to the new class γ. Therefore, it is possible to further assume that those similarities can be captured in feature space by a distribution $p(\boldsymbol{\theta})$. The following experiment tries to uncover whether the knowledge transfered is related to a generic prior or is more category-specific and thus transfers more detailed elements, such as object parts.

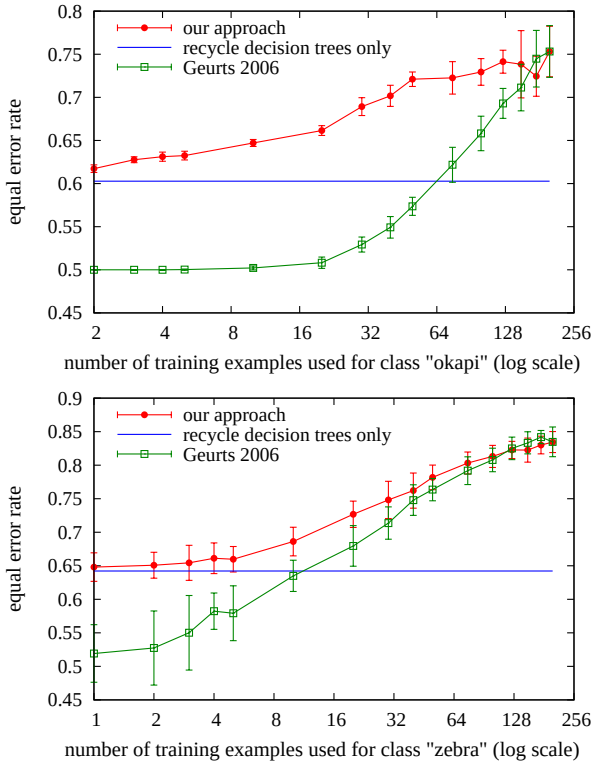

Figure 7: Results of the comparison of our method with the RDF classifier of Geurts et al. (2006) using binary classification tasks.

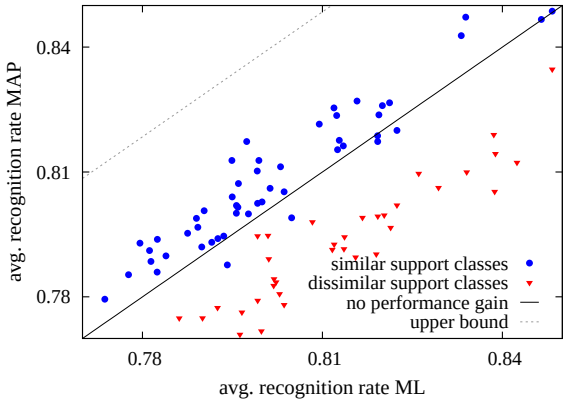

Figure 8: Average recognition rate of the ML approach in comparison to the rate after applying our MAP re-estimation technique. The regularization results in a performance gain only if support classes are (visually) similar to the new class.

To answer this question, an experiment using the letter recognition scenario (Section 7.1) is performed. As a new class with a weak representation of 4 training examples we selected the letter *e* and used two different sets of similar support classes (*a,b,c,d*) and dissimilar support classes (*m,n,w,v,z*). Figure 8 shows a scatter plot of several runs, where each point corresponds to the average recognition rate of a Randomized Decision Forest without (ML estimation) and with our transfer learning method (MAP estimation). All points above the diagonal therefore indicate a clear benefit from prior knowledge. It can be seen that visually dissimilar classes (triangular dots in red color) do not lead to a performance gain and can even decrease the performance.

9.1. Discussion of Experiment 3

Our results clearly show that our transfer learning method learns prior knowledge that is not related to generic prior knowledge. This is an important difference to a lot other approaches which capture more generic prior knowledge. For example in Fei-Fei et al. (2006), Bayesian Interclass Transfer is applied to transfer knowledge between object categories such as: motorbikes, faces, airplanes and wild cats. Therefore, their method seems to use a generic prior of object category images (e.g. size and location of objects are not uniformly distributed). Bart and Ullman (2005) also tested their approach with a large set of various unrelated categories of the Caltech-101 database and showed that the knowledge transfered by their approach, represented by shared image fragments, helped to improve the recognition performance. In general the use of generic prior knowledge has its own tradition and motivation, especially in the context of natural image statistics (Torralba and Oliva, 2003). In our opinion the use of category-specific in addition to generic priors is essential to capture available knowledge as much as possible and thus allows efficient learning with few examples similar to the development of the human visual system.

10. Conclusion

We argue that learning with few examples can benefit from incorporating prior knowledge of related classes (interclass transfer paradigm). Therefore, we proposed to reuse (transfer) the discriminative structure of a Randomized Decision Forest and apply a subsequent maximum a posteriori estimation of leaf probabilities in each tree. This Bayesian formulation allows us to infer knowledge as a prior distribution obtained from related classes and can be seen as a regularization technique. The method is able to exploit interclass relationships to support learning of a class with few training examples.

Experiments on several public datasets showed a significant performance gain in dealing with a weak training representation. In contrast to other work (Fei-Fei et al., 2006), transfer learning of Randomized Decision

Trees is applicable for binary and even for multiclass classification, where information is transferred within the task. An additional experiment validated that the transferred prior information captures (visual) similarities of related classes unlike a generic prior.

11. Further Work

Regularization with a meaningful prior derived from similar object categories is an interesting research direction. Especially for learning with few training examples, transferring knowledge from similar object categories currently seems to be the only way to handle the underlying ill-posed problems.

Despite the benefits presented in this paper, the proposed method has two drawbacks: the support classes have to be selected manually and the influence of the prior has to be controlled by the variance σ^2 of the underlying distribution. The optimal parameter σ^2 could be found by a typical method for estimating the regularization parameter using the L-curve (Kilmer and O'Leary, 2001). An alternative would be to use cross validation, which is a common tool for all parameter estimation problems within a classification task. Automatically selecting the support classes is more complex. In our case it is yet unknown whether the information of few examples is sufficient to estimate the similarity to other categories that would be useful for regularization.

References

Bart, E., Ullman, S., 2005. Cross-generalization: Learning novel classes from a single example by feature replacement. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05). pp. 672–679.

Bonilla, E., Chai, K. M., Williams, C., 2008. Multi-task gaussian process prediction. In: Advances in Neural Information Processing Systems 20. MIT Press, pp. 153–160.

Breiman, L., October 2001. Random forests. Machine Learning 45 (1), 5–32.

Fei-Fei, L., 2006. Knowledge transfer in learning to recognize visual objects classes. In: Proceedings of the International Conference on Development and Learning (ICDL).

Fei-Fei, L., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4), 594–611.

Fink, M., 2004. Object classification from a single example utilizing class relevance pseudo-metrics. In: Advances in Neural Information Processing Systems. Vol. 17. The MIT Press, pp. 449–456.

Fink, M., Ullman, S., 2008. From aardvark to zorro: A benchmark for mammal image classification. Int. J. Comput. Vision 77 (1-3), 143–156.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Maching Learning 63 (1), 3–42.

Hoiem, D., Rother, C., Winn, J., 2007. 3d layoutcrf for multi-view object class recognition and segmentation. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07). pp. 1–8.

Jones, S. S., Smith, L. B., April-June 1993. The place of perception in children's concepts. Cognitive Development 8, 113–139.

Kilmer, M., O'Leary, D., 2001. Choosing regularization parameters in iterative methods for ill-posed problems. SIAM J. Matrix Anal. Appl 22 (4), 1204–1221.

Lazebnik, S., Schmid, C., Ponce, J., 2004. Semi-local affine parts for object recognition. In: British Machine Vision Conference. Vol. 2. pp. 779–788.

Lazebnik, S., Schmid, C., Ponce, J., 2006. A discriminative framework for texture and object recognition using local image features. In: Ponce, J., Hebert, M., Schmid, C., Zisserman, A. (Eds.), Toward Category-Level Object Recognition. Vol. 4170 of Lecture Notes in Computer Science. Springer, pp. 423–442.

Lee, J. W., Giraud-Carrier, C., Aug. 2007. Transfer learning in decision trees. In: International Joint Conference on Neural Networks (IJCNN) 2007. pp. 726–731.

Lee, S.-I., Chatalbashev, V., Vickrey, D., Koller, D., 2007. Learning a meta-level prior for feature relevance from multiple related tasks. In: ICML '07: Proceedings of the 24th International Conference on Machine Learning. pp. 489–496.

Lepetit, V., Lagger, P., Fua, P., 2005. Randomized trees for real-time keypoint recognition. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05). pp. 775–781.

Marée, R., Geurts, P., Piater, J., Wehenkel, L., June 2005. Random subwindows for robust image classification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. pp. 34–40.

Miller, E. G., Matsakis, N. E., Viola, P. A., 2000. Learning from one example through shared densities on transforms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'00). pp. 464–471.

Moosmann, F., Triggs, B., Jurie, F., 2006. Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems. pp. 985–992.

Platzer, E.-S., Denzler, J., Süsse, H., Nägele, J., Wehking, K.-H., October 2008. Challenging anomaly detection in wire ropes using linear prediction combined with one-class classification. In: Proceedings of the Vision, Modelling, and Visualization Workshop. Konstanz, pp. 343–352.

Quattoni, A., Collins, M., Darrell, T., 2007. Learning visual representations using images with captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07). pp. 1–8.

Rodner, E., Denzler, J., October 2008. Learning with few examples using a constrained gaussian prior on randomized trees. In: Proceedings of the Vision, Modelling, and Visualization Workshop. Konstanz, pp. 159–168.

Rohrbach, M., Stark, M., Szarvas, G., Schiele, B., Gurevych, I., 2010. What helps where - and why? semantic relatedness for knowledge transfer. In: CVPR'10: Proceedings of the Computer Vision and Pattern Recognition Conference.

Shotton, J., Johnson, M., Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08). pp. 1–8.

Stark, M., Goesele, M., Schiele, B., 2009. A shape-based object class model for knowledge transfer. In: Proceedings of the International Conference on Computer Vision (ICCV). pp. 373–380.

Tommasi, T., Caputo, B., 2009. The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories. In: BMVC.

Torralba, A., Murphy, K. P., Freeman, W. T., 2007. Sharing visual fea-

tures for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5), 854–869.

Torralba, A., Oliva, A., 2003. Statistics of natural image categories. Network: Computation in Neural Systems 14 (1), 391–412.

van de Sande, K. E., Gevers, T., Snoek, C. G., 2010. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9), 1582–1596.