Learning with Few Examples by Transferring Feature Relevance

Erik Rodner and Joachim Denzler

Chair for Computer Vision
Friedrich Schiller University of Jena
{Erik.Rodner,Joachim.Denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. The human ability to learn difficult object categories from just a few views is often explained by an extensive use of knowledge from related classes. In this work we study the use of feature relevance as prior information from similar binary classification tasks. An approach is presented which is capable to use this information to increase the recognition performance for learning with few examples on a new binary classification task. Feature relevance probabilities are estimated by a randomized decision forest of a related task and used as a prior distribution in the construction of a new forest. Experiments in an image categorization scenario show a significant performance gain in the case of few training examples.

1 Introduction

What is the minimum number of training examples to build robust classification systems? As a human just a single view of one object instance is sufficient in most cases; a machine with current state-of-the-art methods often needs hundreds and thousands of samples.

One possible reason for this gap could be the inability of current systems to determine relevant features of a large pool of generic features from few examples. Another reason is suggested by psychological studies [1] which argue that a key concept of the human ability to recognize from few examples is mainly the concept of *interclass* or *knowledge transfer*. It states that prior knowledge from previously learned categories is the most important additional information source when learning object models from weak representations or few examples [2].

The goal of this paper is to improve the recognition performance of an object recognition system for the case of few training examples using the idea of both explanations. We argue that transferring the relevance of features from related tasks can be very helpful to increase the generalization performance. Feature relevance can be roughly defined as the usefulness of a feature value to predict the class of an object instance mostly defined in terms of mutual information [3]. To give an illustrative example of our transfer idea, consider the recognition of a new animal class. With the aid of prior knowledge from related animal classes,

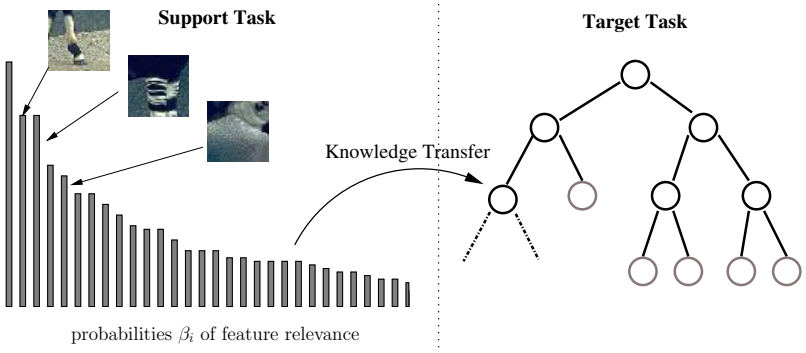

Fig. 1. Illustration of the general principle of our approach: Feature relevance is estimated from a support task and used to regularize the feature selection in the training process (randomized decision forest) of a target task with few training examples. The probabilities and feature visualizations are directly obtained from our image categorization task.

such as the importance of typical body parts like hooves, legs and head, the separation to other categories becomes much easier.

We concentrate on knowledge transfer between two binary classification tasks. A related or support task with a relative large number of training examples and a target task with few training examples is given. We assume that support task and target task share a common set of relevant features. Therefore probabilities of feature relevance for a support task are estimated in a preliminary step. This estimation can use a large number of training examples and thus yields more accurate results than an estimation using just few training examples of the target task. The estimated distribution of feature relevance can then be utilized in the construction process of a randomized decision forest [4]. In contrast to other work [5] which uses a uniform feature distribution, the prior information increases the probability of a relevant feature to be selected for the target task. Fig. 1 provides an overview of this idea.

The remainder of the paper is organized as follows. First of all, we will briefly review related work on object recognition using prior knowledge from related tasks. In Sect. 3 our method is described by first outlining the relationship to Bayesian model averaging. It shows the relationship of our method to the definition of a prior distribution on hypotheses. An estimation technique of feature relevance using a randomized decision forest follows in Sect. 4. Experiments in Sect. 5 show the benefits of our approach in an image categorization task. A summary of our findings and a discussion of future research directions conclude the paper.

2 Related Work

Previous work using the interclass transfer varies significantly in the type of information transferred from related object classes. An intuitive assumption is that similar classes share common geometric intraclass transformations. The *Congealing* approach of Miller et al. [6] therefore tries to estimate those transformations and use them to increase the amount of training data of a target 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 [7, 8]. This leads to methods similar to linear discriminant analysis. Alternatively, Fei-Fei et al. [9] develop a generative framework with maximum-a-posteriori estimation (MAP) of model parameters using a prior distribution estimated from support classes.

The approach of Rodner and Denzler [5] utilizes MAP estimation in a similar sense and re-estimates leaf probabilities of decision trees. As opposed to the approach presented in this paper which builds new decision trees from the scratch, their approach is based on a fixed pre-built decision tree and a fixed set of features not weighted due to their relevancy.

Shared relevant features and class boundaries are exploited in the work of Torralba et al. [10]. They develop a boosting technique that jointly learns several binary classification tasks similar to the combined boosting idea of [11]. Lee et al. [12] transfer feature relevance as a prior distribution on a weight vector in a generalized linear model. Our work is similar to their underlying idea of transfering feature relevance. In contrast prior knowledge in our work is defined using the probability of a feature to be relevant instead of a prior distribution on a specific model parameter. We will show that our approach additionally allows to use a state-of-the-art classifier in form of randomized decision trees [4].

3 Transfer of Feature Relevance (TFR)

Given few training examples a learner tends to overfit and a classification decision is often based on irrelevant or approximately irrelevant features [3]. The goal of our approach is to reduce this overfitting by incorporating a prior distribution $\boldsymbol{\beta}$ on relevant features.

To describe this more precisely, let us first define some simple notations used in the remainder of this paper. Let $\mathcal{T}^S = (\mathbf{x}_i, y_i)_{i=1}^n$ be a set of training examples of a given supporting binary classification task with $y_i \in \{0, 1\}$ and object instances $\mathbf{x}_i \in \mathcal{I}$ (such as an image in an image categorization scenario or an arbitrary multi-dimensional oberservation $\mathcal{I} \subseteq \mathbb{R}^m$). Furthermore let $\mathcal{F}$ be an application-specific set of features $f : \mathcal{I} \rightarrow \mathbb{R}$ that can be calculated on a given object instance. A feature f is said to be relevant for a specific task iff $\exists (\mathbf{x}, y) \in \mathcal{I} \times \{0, 1\} : p(f(\mathbf{x}), y|\mathbf{x}) \neq p(f(\mathbf{x})|\mathbf{x})\ p(y|\mathbf{x})$ and thus retains information about y given $\mathbf{x}$.

Our approach to transfer learning relies on the assumption, that support task and target task share a set $\mathcal{R} \subseteq \mathcal{F}$ of relevant features. We therefore transfer

the probability β_i of a feature f_i to be relevant using the training examples $\mathcal{T}^S$ of the related task:

$$p(\tilde{\mathcal{R}} \mid \mathcal{F}) = p(\tilde{\mathcal{R}} \mid \mathcal{F}, \mathcal{T}^S) = \prod_{f_i \in \tilde{\mathcal{R}}} \underbrace{p(f_i \in \mathcal{R} \mid \mathcal{F}, \mathcal{T}^S)}_{\beta_i} \ . \tag{1}$$

The last reformulation in (1) assumes that the relevance of features is independently distributed. While we delay the estimation of $\boldsymbol{\beta}$ to Sect. 4, the following section shows that $\boldsymbol{\beta}$ can be used as a prior distribution on the set of possible hypotheses or models for the target task. This also highlights that our prior knowledge can be easily integrated in the concept of randomized classifier ensembles and specially the randomized decision tree approach of [4].

3.1 Incorporation of TFR into Randomized Classifier Ensembles

We will now describe the randomized forest approach of Geurts et al. [4] in a theoretical framework related to Bayesian model averaging. This allows to motivate the transfer of feature relevance as a Bayesian approach of defining a prior distribution on models or hypotheses.

The final goal is to estimate the probability of the event Ω that a previously unseen object instance $\mathbf{x}$ belongs to class $y = 1$ conditioned on the set of few training examples of the target task $\mathcal{T}^T$ and the set of all possible features $\mathcal{F}$. As a classification model we will use an ensemble of base models h (in our case non-randomized single decision trees) in the following sense:

$$p(\Omega \mid \mathbf{x}, \mathcal{T}^T, \mathcal{F}) = \int_h p(\Omega \mid \mathbf{x}, h)\ p(h \mid \mathcal{T}^T, \mathcal{F})\ \mathrm{d}h \ . \tag{2}$$

The model h is often assumed to be deterministic for a given training and feature set, but there are multiple ways to sample from those sets and thus generate multiple models. One idea is the concept of bagging [13] which uses random subsets of the training data. As proposed in [13] and [4] another possibility is to use random subsets of all features. This approach can be regarded as Bayesian model averaging and reflects our uncertainty about the set $\mathcal{R}$ of relevant features:

$$p(h \mid \mathcal{T}^T, \mathcal{F}) = \sum_{\tilde{\mathcal{R}} \subseteq \mathcal{F}} p(h \mid \mathcal{T}^T, \tilde{\mathcal{R}})\ p(\tilde{\mathcal{R}} \mid \mathcal{F}) \ . \tag{3}$$

The distribution $p(\tilde{\mathcal{R}} \mid \mathcal{F})$ describes the probability that $\tilde{\mathcal{R}}$ is the set of relevant features. A base model h is deterministic given a training set and the set of relevant features: $p(h \mid \mathcal{T}^T, \tilde{\mathcal{R}}) = \delta(h - h(\mathcal{T}^T, \tilde{\mathcal{R}}))$. Combining all equations yields the final classification model:

$$p(\Omega \mid \mathbf{x}, \mathcal{T}^T, \mathcal{F}) = \sum_{\tilde{\mathcal{R}} \subseteq \mathcal{F}} p(\Omega \mid \mathbf{x}, h(\mathcal{T}^T, \tilde{\mathcal{R}}))\ p(\tilde{\mathcal{R}} \mid \mathcal{F}) \ . \tag{4}$$

This sum can not be computed efficiently for large feature spaces, therefore we can approximate it by simple Monte Carlo estimation:

$$p(\Omega \mid \mathbf{x}, \mathcal{T}^T, \mathcal{F}) = \frac{1}{M} \sum_{i=1}^{M} p(\Omega \mid \mathbf{x}, h(\mathcal{T}^T, \mathcal{R}^i)) \ . \tag{5}$$

Feature subsets $\mathcal{R}^i$ are sampled from $p(\tilde{\mathcal{R}} \mid \mathcal{F})$. This distribution is often assumed to be uniform and samples of only a fixed number of training instances $|\tilde{\mathcal{R}}| = m$ are used [4]. This assumes that we have a prior estimate of $|\mathcal{R}|$ or the integral of (4) can be nevertheless approximated by a subspace of the power set of $\mathcal{F}$.

We now apply the idea of our transfer learning technique that was described at the beginning of Sect. 3. Instead of using a uniform distribution $p(\tilde{\mathcal{R}} \mid \mathcal{F})$ one can use the probabilities $\beta_i = p(f_i \in R \mid \mathcal{F}, \mathcal{T}^S)$ obtained from the related class. This prior information reduces the uncertainty of the learner about the optimal set of relevant features.

In the following section we will briefly outline the special characteristics of randomized decision trees and the connection to the previous description of general randomized ensembles.

3.2 Randomized Decision Trees

As we use randomized decision trees [4] our base models are decision tree classifiers. Those classifiers are binary trees with two types of nodes. Each inner node represents a weak classifier, a feature f and a threshold, which defines a hyperplane in feature space and thus determines the traversal of a new example in the tree. The traversal ends in a leaf node n. Each of those nodes is associated with a posterior distribution $p(\Omega \mid n)$, which is an estimation of the probability of the object class given that this specific leaf is reached. Building a tree is done by iteratively splitting the training set with the most informative weak classifier. The selection of a weak classifier is done by choosing the weak classifier with the highest gain in information from a random fraction of features $\mathcal{R}$ and possible thresholds.

Note that in contrast to the theoretical explanation in Sect. 3.1, the selection of a random subset of features is performed in each node rather than a single time for the whole decision tree. This fact is also highlighted by the illustration in Fig. 1. Relevant features are sampled from the distribution $\boldsymbol{\beta}$ in each split node during the training process.

Using just few training examples leads to decision trees of small depth. Due to the model averaging technique described in the previous section and by using the idea of bagging [13] they still allow to build robust classifiers.

4 Estimating Feature Relevance using Randomized Decision Forests

As pointed out by Rogers and Gunn [14], the use of ensembles of decision trees allows to provide robust estimates of feature relevance that also incorporates

dependence between features. Our technique is similar to their method which uses a modified average mutual information between a feature and the class variable y in each inner node.

The first step to estimate underlying feature relevance of the supporting task is the training of a randomized decision forest with all training examples. Afterwards we count the number of times c_i a feature f_i is used in a split node. A feature with a high occurrence c_i is likely to be relevant for this task. We did not directly use the mutual information associated with a split node because our goal is to estimate a well defined distribution rather than a relevance ranking.

To obtain the final vector $\boldsymbol{\beta}$ of feature relevance, we use maximum-a-posteriori estimation:

$$\boldsymbol{\beta}^{MAP} = \arg\max_{\boldsymbol{\beta}} p(\mathcal{T}^S \mid \boldsymbol{\beta})\; p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \arg\max_{\boldsymbol{\beta}} \left(\prod_i \beta_i^{c_i} \right)\; p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \qquad (6)$$

with $\boldsymbol{\alpha}$ being the hyperparameter of a Dirichlet prior $p(\boldsymbol{\beta})$ and $\forall i : \; \boldsymbol{\alpha}_i = \alpha$. Without this prior distribution, the optimal $\boldsymbol{\beta}$ is the normalized vector $\mathbf{c}$ of all counts. The prior distribution can be thought of as a smoothing term, that prevents zero probability of relevance for some features. This is theoretically important if there is a feature f_i that is completely irrelevant for the supporting tasks, but with discriminative power in the target class.

In Sect. 5.2 we evaluate the influence of parameter α. Surprisingly it turns out that in our experimental setting a flat prior distribution ($\alpha = 1$) is sufficient.

5 Experiments

We experimentally evaluated our approach to illustrate the benefits of transferring feature relevance from related tasks. In the following we empirically validate the following hypotheses:

1. Transferring feature relevance (TFR) from related tasks helps to improve recognition performance in the case of few examples.
2. The benefit is most prevalent, if the supporting task is visually similar to the target task.
3. A smoothing of feature relevance is not necessary ($\alpha = 1$).

We use image data from the Caltech 101 dataset [9] to show the applicability to image categorization tasks. Three classes were used to conduct binary classification tasks: Okapi, Gerenuk and Chair vs. the Caltech background class with 200 training images (cf. Fig. 2).

To use the transfer of feature relevance supporting task and target task should use a common feature representation. In our experiments we used a bag-of-features representation as described in Sect. 5.1. Therefore a bag-of-features codebook of the supporting class is used to calculate features of the target class.

Measuring the performance of the binary classification tasks is done by the area under the ROC curve (AUC). Unless additionally specified we obtain a

Fig. 2. Example images of all three classes from the Caltech 101 database [9] which are used for evaluation: Okapi, Gerenuk and Chair.

statistically meaningful estimate of the performance by calculating the average of 10 runs with a random subset of the training data. Due to the behavior of randomized decision trees for each of those subsets the classifier is trained and tested 50 times and the performance values are also averaged. This results in 500 runs in total which produce the final AUC value for a specific setting.

5.1 Feature Extraction

A standard approach to image categorization is the bag-of-features idea. A quantization of local features, which is often called a codebook, is computed at the time of training. For each image a histogram can then be calculated which counts for each codebook entry the number of matching local features.

The method of [15], which utilizes a randomized decision forest as a clustering mechanism, is used to construct the codebook This codebook generation procedure showed superior results compared to standard k-Means in all experiments. It also allows to create large codebooks in a few seconds on a standard PC. Local features are extracted for each image by dense sampling of feature points with a horizontal/vertical pixel spacing of 10 pixels. Descriptors are calculated using Opponent-SIFT [16].

5.2 Dirichlet Parameter

In a first experiment we evaluate the influence of the generic prior distribution in equation (6). This data-independent prior distribution serves as a smoothing term for the estimation of relevant features. We build a randomized decision forest using a fixed set of 30 examples of the Okapi class and 200 examples of the background class. From this randomized decision forest we estimate feature relevance as described in Sect. 4 with a varying Dirichlet parameter α. These estimates are used afterwards to classify a set of one training example from the Gerenuk class and the same background images used before. Average performance values and standard deviation of 50 runs are illustrated in Fig. 3(a).

It can be seen that with an increasing value of α the performance drops and the optimal value remains at $\alpha = 1$. For this reason we fix α to this value, which corresponds to maximum likelihood estimation of $\boldsymbol{\beta}$. This highlights that the complete removal of features which are irrelevant for the supporting task ($p(f_i \in \mathcal{R}) = 0$) is beneficial. This may be not the case in situations with a

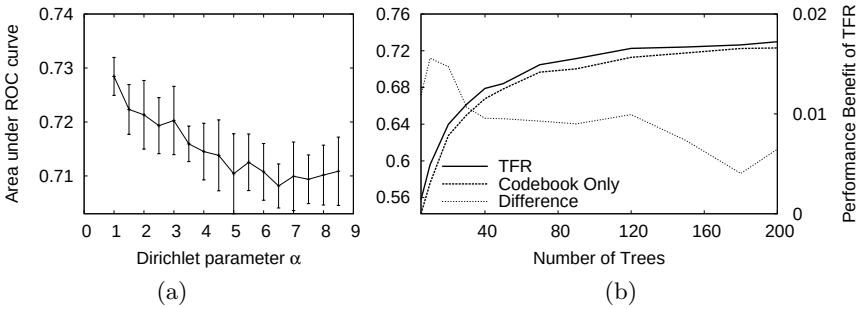

Fig. 3. (a) Evaluation of the hyper-parameter α of the Dirichlet distribution, which is used to smooth the probabilities of feature relevance. (b) Influence of the number of decision trees in the forest.

smaller feature set and features, that are completely irrelevant for the supporting task but essential for a target task.

5.3 Influence of the Ensemble Size

We analyzed the influence of the number of base learners in the ensemble. The the same experiment as in Sect. 5.2 is performed with a varying size of the ensemble. The results are illustrated in Fig. 3(b).

The performance obviously increases with an increasing number of base learners in the ensemble which was also proven theoretically and empirically in [13]. Another interesting effect is that the performance benefit of transferring feature relevance is most prevalent when a small number of base learners are used. For all other experiments we utilized randomized decision forests with 200 trees.

5.4 TFR Improves Learning with Few Examples

We analyzed the performance of the binary classification task Okapi class vs. background class with different types of knowledge transfer: transfer of feature relevance using the support classes Gerenuk or Chair; using only the codebook of the supporting class; no knowledge transfer of other tasks at all. Note that no knowledge transfer means that a codebook is generated only from training examples of the target task. Fig. 4(a) illustrates the resulting recognition rates. Additionally Fig. 4(b) shows the same results for the class Gerenuk with prior information learned also from the Okapi task.

At first it can be seen that transferring feature relevance from related tasks really improves the recognition performance compared to a method which uses no knowledge transfer at all. This performance benefit is most prevalent with a visually similar class such as the related animal class. Using prior knowledge the chair class is sometimes also beneficial. It is most likely that this is due to the learning of natural generic prior knowledge, which also showed in other work to improve recognition performance [9].

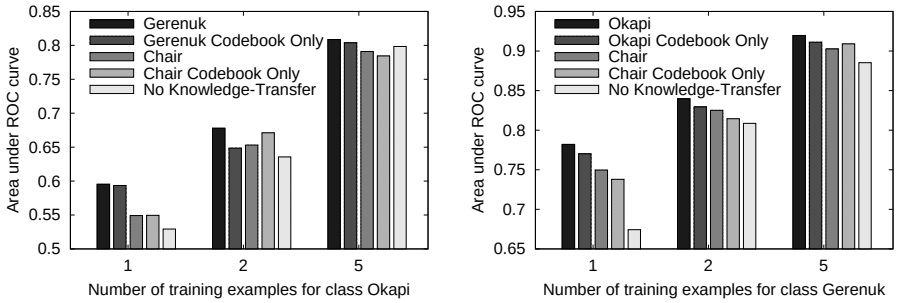

Fig. 4. Experiments with the target task 4(a) "Okapi" and 4(b) "Gerenuk" vs. background and several types of support tasks with a varying number of training examples of the target task.

Transferring only the codebook from the supporting task also increases the performance. The difference between TFR and this method in Fig. 4(b) for one training example might seem to be minor at first glance and insignificant due to a standard deviation of about 1% in the previous experiment (Fig. 3(a)). But using a paired t-test and the corresponding average results of all 10 training and test runs, we are able to show significance with a level of $p < 0.003$.

6 Conclusion

We presented a classification approach that transfers knowledge from related classification tasks to improve the recognition performance on a task with few training examples. The key concept of our method is the transfer of feature relevance. We use probabilities of feature relevance, which are estimated using a randomized decision forest of a related task. Those probabilities form a distribution that is used to select a random subset of features during the building process of a randomized decision forest for the target class with few examples. The relationship of our method to Bayesian model averaging was outlined. It shows that the technique indirectly uses a prior distribution of hypotheses to regularize the training of a target classification task.

Experiments on an image categorization task show a significant performance gain. This performance benefit is most striking if the supporting binary classification task is visually related to the task with few training examples.

7 Further Work

The presented method can be applied to arbitrary machine learning problems and is not restricted to generic image categorization. For this reason we plan to apply the classification technique to other application areas, such as object

localization. An interesting open question is, whether the general idea to transfer feature relevance can be applied to other classifier techniques such as support vector machines. Additionally our method of feature relevance estimation should be compared to other methods of feature selection.

References

1. Jones, S.S., Smith, L.B.: The place of perception in children's concepts. Cognitive Development **8** (April-June 1993) 113–139
2. Fei-Fei, L.: Knowledge transfer in learning to recognize visual objects classes. In: Proceedings of the International Conference on Development and Learning (ICDL). (2006)
3. Guyon, I., Gunn, S., an d Lotfi A. Zadeh, M.N.: Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing). Springer (2006)
4. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Maching Learning **63**(1) (2006) 3–42
5. Rodner, E., Denzler, J.: Learning with few examples using a constrained gaussian prior on randomized trees. In: Proceedings of the Vision, Modelling, and Visualization Workshop, Konstanz (October 2008) 159–168
6. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: IEEE Conference on Computer Vision and Pattern Recognition. (2000) 464–471
7. Fink, M.: Object classification from a single example utilizing class relevance pseudo-metrics. In: Advances in Neural Information Processing Systems. Volume 17., The MIT Press (2004) 449–456
8. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
9. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(4) (2006) 594–611
10. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(5) (2007) 854–869
11. Levi, K., Fink, M., Weiss, Y.: Learning from a small number of training examples by exploiting object categories. In: CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop. Volume 6. (2004) 96–104
12. Lee, S.I., Chatalbashev, V., Vickrey, D., Koller, D.: Learning a meta-level prior for feature relevance from multiple related tasks. In: ICML '07: Proceedings of the 24th International Conference on Machine Learning. (2007) 489–496
13. Breiman, L.: Random forests. Machine Learning **45**(1) (October 2001) 5–32
14. Rogers, J., Gunn, S.R.: Identifying feature relevance using a random forest. In: Subspace, Latent Structure and Feature Selection, Statistical and Optimization, Perspectives Workshop. (2005) 173–184
15. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems. (2006) 985–992
16. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluation of color descriptors for object and scene recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8