

Enhanced Representation and Multi-Task Learning for Image Annotation

Alexander Binder^{a,c,1,*}, Wojciech Samek^{a,c}, Klaus-Robert Müller^{a,1}, Motoaki Kawanabe^{b,c,1}

^a*Machine Learning Group, Berlin Institute of Technology (TU Berlin)
Franklinstr. 28/29, 10587 Berlin, Germany*

^b*ATR Brain Information Communication Research Laboratory Group
2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288, Japan*

^c*Fraunhofer Institute FIRST
Kekuléstr. 7, 12489 Berlin, Germany*

Abstract

In this paper we evaluate biased random sampling as image representation for bag of words models in combination with between class information transfer via output kernel-based multi-task learning using the ImageCLEF PhotoAnnotation dataset. We apply the mutual information measure for measuring correlation between kernels and labels. Biased random sampling improves ranking performance of classifiers and mutual information between kernels and labels. Output kernel multi-task learning (MTL)² permits asymmetric information transfer between tasks and scales to training sets of several thousand images. The learned contributions of source tasks to target tasks

*Corresponding author

Email addresses: alexander.binder@tu-berlin.de (Alexander Binder),
wojciech.samek@campus.tu-berlin.de (Wojciech Samek), kawanabe@atr.jp (Motoaki Kawanabe)

URL: <http://www.ml.tu-berlin.de> (Klaus-Robert Müller)

¹research supported by the THESEUS program <http://www.theseus-programm.de/>

²multi-task learning

are shown to be semantically consistent.

Our best visual result which used the MTL method was ranked first according to mean average precision (mAP) within the purely visual submissions in the ImageCLEF 2011 PhotoAnnotation Challenge. Our multi-modal submission achieved the first rank by mAP among all submissions in the same competition.

Keywords: Image Ranking, Image Classification, Multiple Kernel Learning, Multi Task Learning, Bag-of-Words Representation, Biased Random Sampling, ImageCLEF, Mutual Information

1. Introduction

Learning machines have been successfully employed in a variety of scientific fields such as Chemistry, Physics, Neuroscience and have become standard techniques in industrial data analysis. A particularly hard learning task is machine computer vision where the data is highly complex and even seemingly simple questions such as image annotation are easy for humans, but extraordinarily hard for a machine. In this paper we take a statistical approach to image annotation and show novel algorithmic contributions that help to push the boundaries of this problem.

We focus on two aspects, one at the early stage of image annotation and one at a late stage. Firstly, we consider biased random sampling for the selection of local features in a bag of words (BoW) ³ model. Secondly, we attempt to transfer information between semantic concepts by computing

³bag of words

kernels from classifier outputs and combining them using non-sparse multiple kernel learning (MKL)⁴.

We evaluate these contributions on the 99 semantic concepts of the ImageCLEF 2011 PhotoAnnotation Dataset [1] which itself was created from the MIRFLICKR-1M dataset [2]. The methods presented here have been used in the ImageCLEF 2011 PhotoAnnotation Challenge on 10000 images with undisclosed concept labels. Our visual submissions using biased random sampling and multi-task information transfer and our multi-modal submission also using biased random sampling achieved both the best results by the mean Average Precision (mAP) measure among the purely visual and multi-modal submissions, respectively.⁵

The introduction concludes by discussing related work in Section 1.1 and the ImageCLEF Dataset in Section 1.2. Section 2 gives an overview over Bag-of-Word features and our particular setup. The questions addressed in our experiments and evaluated here are

- *Local feature computation*: What impact does biased random sampling have on feature properties and the ranking quality for a set of semantic concepts? See Section 3.
- *Classifier combinations*: Can we transfer information between classes to improve ranking performance? See section 4.

⁴multiple kernel learning

⁵Submissions by other groups like BPACAD (Hungarian Academy of Sciences, Budapest, Hungary), CAEN (University of Caen, France), ISIS (ISIS Group of the University of Amsterdam, Netherlands) and LIRIS (LIRIS Group of the University of Lyon, France) were ranked closely.

1.1. Related Work

1.1.1. Biased Random Sampling

The detection of salient regions has attracted research efforts over the past years. The human ability to find unspecific but striking content in a short glimpse or specific objects in a prolonged eyeballing session serves as the leading motivation. Saliency algorithms have been developed for various scales: considering differences compared to sets of images, global differences within a single image or local differences within a single image. center surround differences over multiple scales and feature cues are used in [3] in order to find parts of an image which are salient which belongs to the last kind in the above consideration. Another approach in [4] discusses saliency based on color edges using a multi-scale approach which resorts to first two ideas in the above consideration. Saliency in [4] is linked to inverse frequency of occurrence and equally frequent edges are assigned the same saliency value. This in turn is based on [5]. Saliency based on Shannon entropy has been discussed in [6]. Outlier analysis based on clustering as a tool for defining saliency has been used in [7].

The link between saliency maps and object categorization is introduced in [8] by learning to sample subwindows of an image. Biased random sampling for local features of a BoW model has been used in [9] by using a top-down object class prior map and a bottom up visual saliency model inspired from human visual processing.

1.1.2. Information Transfer between Classes

Relations between concepts, e.g. co-occurrence and exclusiveness, spatial and functional dependencies or common features and visual similarity,

provide an important source of information which is efficiently utilized by humans. For instance, knowing that an image contains chairs, knives and forks strongly suggests that there may be a table in the image, simply because these items are functionally related and often co-occur in images. Therefore transferring information about other classes may improve ranking performance.

Recently, several approaches were proposed to incorporate such information into the image annotation task. One promising approach to transfer knowledge from some classes to others is to share information between object classes on the level of individual features [10, 11, 12, 13, 14] or entire object class models [15, 16]. Other algorithms like [17, 18, 19] were proposed for zero-shot learning in a large-scale setting, i.e. to learn concepts with very little or without training examples by transferring information from other classes. Another way to transfer information between classes is to incorporate the relations into the learning step. There exists also a large literature on Multi-Task Learning (MTL) methods which aim at achieving better performance by learning multiple tasks simultaneously. One principled approach [20] is to define a kernel between tasks and treat the multi-task problem among the lines of structured prediction [21], i.e. to represent relations between tasks by a kernel matrix and perform multi-task learning by using a tensor product of the feature and task kernel as input for SVM. Although this method is theoretically attractive, it has some drawbacks which reduce its applicability in practice. For example it requires to define a kernel between tasks even when one has no clear prior knowledge about similarities between tasks and the interactions between tasks are symmetric. This implies that

bad performing tasks may spoil the performance of better performing ones. Finally, structured prediction leads to one big optimization problem working typically on a kernel of squared size number of tasks times the number of samples. This makes the application on training data with many training samples and many tasks challenging in practice. Another kernel-based approach for multi-task learning [22] uses projections of external cues to weight support vectors depending on the sample to be classified. Several other Multi-Task Learning methods have been proposed like [23] where the authors used Gaussian Processes for learning multiple-tasks simultaneously.

1.2. The Dataset

We evaluated our methods on the data from the annotation task of the ImageCLEF2011 Photo Annotation Challenge [1] because of its diverse large set of semantic concepts and challenging images.

The challenge task required the annotation of 10000 images in the provided test corpus according to the 99 pre-defined semantic concepts. Note that this year’s ImageCLEF photography-based challenge provides additionally a second challenging competition [1], a concept-based retrieval task. In the following we will focus on the firstly mentioned annotation task over the 10000 images. This dataset comes with 8000 annotated training images. The annotation belongs to the multi-label category, i.e. each image may belong to many concepts simultaneously. The data in the corpus is a subset of the MIRFLICKR-1M dataset [2].

The ImageCLEF photo corpus is challenging due to its high variance in appearance. The images are real world photographs with varying lighting conditions and image qualities stemming from many camera models. Visual

cues and objects have varying scales and positions. This constitutes a difference for example to the seminal Caltech101 dataset [24] which is still popular in research today. The images in it are centered and show objects of the same scale up to the point that for many semantic classes the class label may be inferred from the average image of all images of one class.

Another source of complexity is the heterogeneity of semantic concepts. It contains classes based on concrete tangible objects such as *female*, *cat* and *vehicle* as well as more abstract classes such as *Technical*, *Boring*, *Still Life* or *Aesthetic_Impression*. Some of these classes could be approached using object detectors finding bounding boxes. Others may not. This constitutes a difference to the object-based concepts used in the Pascal VOC Challenge.

Finally, many of the concepts such as *Scary*, *Euphoric*, *Still Life* are highly subjective. This implies from a statistical viewpoint an unknown but varying amount of label noise.

2. Pipeline Description

All our experiments are based on discriminatively trained classifiers over kernels [25] using Bag-of-Words (BoW) features. The generic Bag-of-Words pipeline can be subsumed by Figure 1. The coarse layout of our approach is influenced by the works of the Xerox group on Bag-of-Words in Computer Vision [26], the challenge submissions by INRIA groups [27] and the works on color descriptors by the University of Amsterdam [28]. It is based on computing many BoW features while varying, firstly, the sets of color channels of an image used for the computation of local features, and secondly the spatial tilings of an image for computation of multiple concatenated BoW vectors.

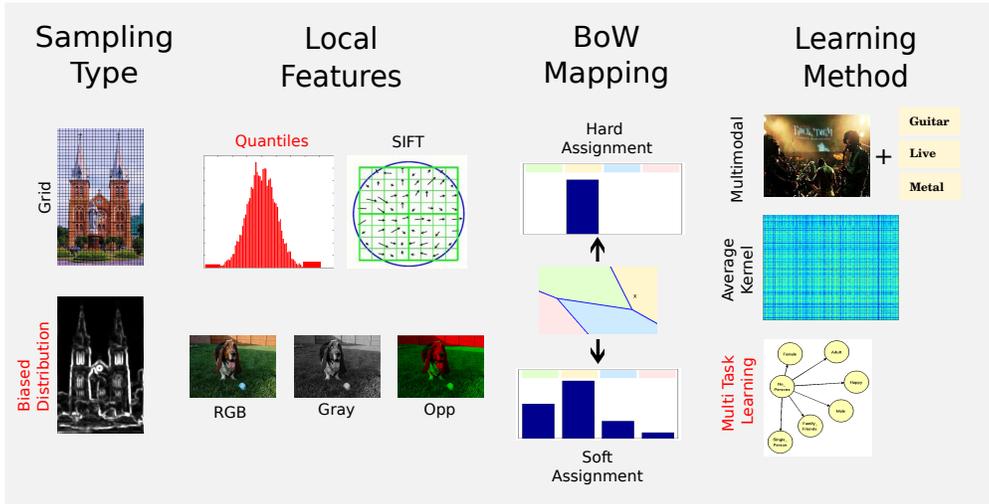


Figure 1: Bag-of-Words pipeline.

Table 1 shows the computed BoW features. Information about the sampling methods is given in Section 3. We used color channel combinations red-green-blue (RGB), grey (Gr), grey-opponentcolor1-opponentcolor2 (Opp) and a grey-value normalized version of the last combination (N-Opp in Table 1). For understanding of these color descriptors see [28]. For SIFT feature computation we used the software underlying [28] available at www.colordescriptors.com. The SIFT features are computed without orientation invariance and the grid has step size six. Another feature were nine-dimensional vectors of quantiles corresponding to 10% to 90% percentiles over local pixel color distributions denoted as *Quantiles* in Table 1.

For mapping the local features onto visual words we used either the usual hard mapping denoted as Hard in Table 1 or rank-based soft mapping which works as follows. The mapping of local features onto visual words is shown in Figure 1. See [29] for an introduction to soft mappings. Be $RK_a(l)$ the rank

Sampling	Local	Color	BoW	No. of	BoW
	Feature	Channels	Mapping	Features	Dimensions
grid	Quantiles	RGB, Opp, Gr, weighted Hue	Rank	12	900
grid	SIFT	RGB, Opp, Gr, N-Opp	Hard	12	4000
grid	SIFT	RGB, Opp, Gr, N-Opp	Rank	12	4000
bias1	SIFT	RGB, Opp, Gr	Rank	9	4000
bias2	SIFT	RGB, Opp, Gr, N-Opp, N-RGB	Rank	15	4000
bias3	SIFT	RGB, Opp	Rank	6	4000
bias4	SIFT	RGB, Opp, Gr	Rank	9	4000

Table 1: Bag-of-Words Feature Sets. See text for explanation.

of the distances between the local feature l and the visual word corresponding to BoW dimension d , sorted in increasing order along all visual words. Then the BoW mapping m_d for dimension d is defined as ⁶:

$$m_d(l) = \begin{cases} 2.4^{-RK_d(l)} & \text{if } RK_d(l) \leq 8 \\ 0 & \text{else} \end{cases} \quad (1)$$

For all visual Bag-of-Words features we computed χ^2 -Kernels. The kernel width has been set as the mean of the χ^2 -distances on the training data. All kernels have been normalized to standard deviation one as this allows to

⁶for understanding of the constant: $2.4^8 \approx 1000$

Submission	Modality	mAP	Submission	Modality	mAP
BPACAD 3	T	34.56	TUBFI 1	V	38.27
IDMT 1	T	32.57	TUBFI 2	V	37.15
MLKD 1	T	32.56	TUBFI 5	V	38.33
TUBFI 3	V+T	44.34	TUBFI 4	V	38.85
LIRIS 5	V+T	43.70	CAEN 2	V	38.24
BPACAD 5	V+T	43.63	ISIS 3	V	37.52

Table 2: Results by mAP on test data with undisclosed labels for the best three and our own submissions (10000 images, 99 concepts). T:Textual task, V:Visual, V+T:Multimodal. See footnote in Section 1 for the full names of the groups mentioned here.

search for optimal regularization constants in the vicinity of 1 in practice.

For the detailed results of all submissions we refer to the overview given in [1]. A small excerpt can be seen in Table 2.

We will discuss the random biased sampling and the output kernel Multi-Task Learning (MTL) in more detail as the improvements from these techniques are generic and have value beyond the submissions to ImageCLEF.

3. Biased random Sampling

Humans are able to capture the most salient regions an image within a short time. The detection of salient regions has attracted research interest over the past years.

The application of salient region detectors to object categorization has been made in [8] for the selection of image subwindows. [9] uses saliency

detection for BoW models. Random biased sampling in the context of BoW models is based on drawing centers of computation regions for local features from a probability distribution over pixels of the image.

However the BoW model seems to be different from many models of human visual processing. Therefore we argue that the saliency detectors for usage with the BoW model need to be adapted to its specific properties, particularly taking the local features into consideration. Given that the best local features for bag of word models are gradient-based ones like SIFT [30] or SURF [31] we aim at extracting extract gradient-based features centered on gradient-rich regions.

Furthermore it might be useful to adapt the scale of the detectors to the scale of the local features whereas the usual way is based on the opposite idea of extracting local features on a scale predicted by the saliency detector. Since the saliency detectors are not adapted to BoW models their scale predictions might be not optimal for usage with BoW features. This constitutes a difference to approaches using across scale differences and multi-scale analysis.

Finally, given the variety of concepts in the ImageCLEF 2011 PhotoAnnotation dataset which contains global impression concepts like *Calm* or background concepts like *Sky* it seems reasonable not to suppress feature extraction from larger parts of the image completely as done with some salient region approaches. For the same reason we refrained from early decisions like object detection as in [5].

[9] uses a bottom-up attention bias model and top-down object class prior maps. The latter made sense for the data sets used in the original

publication but seemed to have no good justification for many of the more abstract ImageCLEF concepts which do not rely on the presence of a single object such as e.g. *Natural*, *PartyLife*. Therefore we also used only the model for an attention based bias [3] named *bias3* in the following and introduced gradient-based biases for probabilistic sampling of local features.

Such probabilistic sampling approaches offer two potential advantages over Harris Laplace detectors: Firstly, we get key points located on edges rather than corners. A motivating example can be seen in Figure 3 – the bridge contains corner points but the essential structures are lines. Similar examples are smooth borders of buildings, borders between mountains and sky, or simply a part of circle with low curvature. In that sense our approach can be compared to using the absolute value of the Harris corner measure.

Secondly, we did adjust the number of local features to be extracted per image as a function of the image size instead of using the typical corner detection thresholds. The reference is half of the number of local features extracted by grid sampling, in our case six pixels. This comes from the idea that some images e.g. showing large portions of sky can be more smooth in general which may lead to a very low number of extracted key points. However too sparse sampling of key points may hurt the recognition and ranking performance substantially as shown in [32] and also supported by the good results of per pixel dense sampling in [33]. Given such results we think that forcing the number of key points to be extracted to be proportional to the image size rather than fixed thresholds is the second key to the good performance of our key points.

The usage of random biased sampling over thresholded keypoint detectors

has a further qualitative justification. The BoW feature can be represented as an expectation over the extracted local features. Consider the BoW mapping over local features l computed from regions defined by points p on an image drawn from a probability distribution P . Be m the vector-valued mapping of local features onto visual words. The mapping of local features onto visual words is shown in Figure 1. One example is the hard mapping which adds for each local feature a constant to the nearest visual word in the resulting BoW feature. Then the BoW feature v can be represented as in equation 2. Biased random sampling corresponds to a choice of a probability measure P in this framework.

$$v = \int_p m(l(p))dP(p) . \quad (2)$$

3.1. Biased Sampling Methods

In the following we describe four detectors *bias1* to *bias4* which have been used in our experiments. Their probability maps for one composite image depicting many natural scenes in one are shown in Figure 2.

bias3 is a simplified version of an attention based detector [3]. However this detector requires to set scale parameters for across scale differences. The highly varying scales of motifs in the ImageCLEF dataset makes it difficult to find a globally optimal set of scales without expensive optimizations. This inspired us to try detectors which depend to a lesser degree on scale parameters:

bias1 computes an average of gradient responses over pixel-wise images of the following color channels: grey, red minus green, green minus blue and



Figure 2: An image and its probability maps from biased random sampling. Upper Left: *bias3*, Upper Right: *bias1*, Lower Left: *bias2*, Lower Right: *bias4*.

blue minus red. The probability map is given as

$$P(x, y) \propto \left(\exp \left(\sum_c B_1 \circ \|\nabla I_c(x, y)\|_2 \right) - 1 \right) \exp(B_2 \circ I_g(x, y)) \quad (3)$$

where $I_g(x, y)$ is the grey channel of the original image at pixel (x, y) , $I_c(x, y)$ is one of the above color difference channels, B_σ is a Gaussian blur operator using width σ . ∇ denotes an image gradient operator. The product of exponentials achieves a weak weighting using image brightness $I_g(x, y)$ and a stronger weighting based on the gradients over the grey and the color difference channels $I_c(x, y)$.

bias2 is like *bias1* except for dropping the grey channel and using a maximum instead of the sum over color channels. The probability map is given as using the same definitions as for equation 3

$$P(x, y) \propto \left(\exp \left(\max_c B_1 \circ \|\nabla I_c(x, y)\|_2 \right) - 1 \right) \exp(B_2 \circ I_g(x, y)) \quad (4)$$

Thus it will fail on grey images but detects strong local color variations. On the other hand such differences between RGB color channels are more prominent on bright regions. This allows to use features over normalized color channels more safely on color images.

bias4 takes the same set of color channels as the underlying SIFT descriptor and computes the entropy of the gradient orientation histogram on the same support as the SIFT descriptor would use. We normalize the negative entropy to lie in $[0, 1]$ and take the pixel-wise product with the gradient norm. These products are averaged over all used color channels and scales.

$$P(x, y) \propto \sum_s \sum_c (1 - E(\text{supp}_s[I_c(x, y)])) (B_1 \circ \|\nabla I_c(x, y)\|_2) \quad (5)$$

Let $\text{supp}_s[I_c(x, y)]$ be the region mask to be used for the computation of a local feature at scale s centered on pixel (x, y) . The region mask is applied to the color channel I_c . E is the Shannon entropy normalized to $[0, 1]$. Regions with low entropy are preferred in the probability map used for biased random sampling. This detector is adapted more closely to the SIFT feature by computing a score on its support and color channel. The question behind this detector is whether the focus on regions with low entropies in gradient orientation distributions constitutes an advantage. The usage of the entropy as measure is reminiscent of the salient region detector from [6] but it is adapted to the color channels and scale of the local feature instead of using differences across scales as in [6]. A comparison of the extracted key points to a Harris Laplace detector is shown in Figure 3.

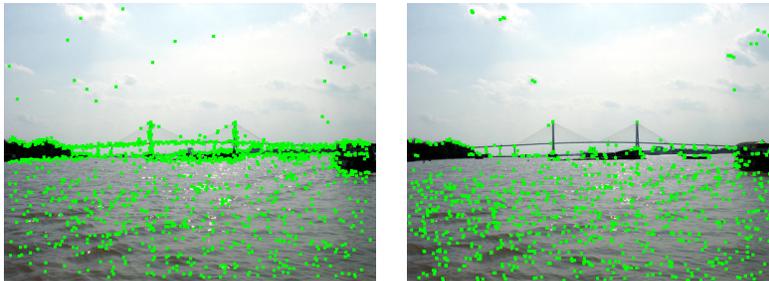


Figure 3: The essential structures of the bridge are lines rather than corners. Left: *bias4* key points. Right: Harris Laplace using the same number of key points.

3.2. Insights from Random Biased Sampling

Here we evaluate the performance of our BoW features which were based on biased random sampling.

sampling type	<i>grid</i>	<i>bias1</i>	<i>bias4</i>
mAP	36.71 ± 6.469	36.75 ± 6.686	36.85 ± 6.696
sampling type	<i>bias2</i>	<i>bias3</i>	all
mAP	36.49 ± 6.471	35.04 ± 6.423	38.31 ± 6.792

Table 3: Mean Average Precisions for varying sampling types, computed over averages kernel based on six kernels: RGB- and Opponent-SIFT and all three spatial tilings. Results were computed on training data via 12-fold cross-validation.

3.2.1. Ranking Performance Evaluation

In Table 3 we see that biased sampling is more or less equally strong in mAP ranking performance, however, it uses at most half of the local features as dense sampling; this implies that it is at least two times faster during BoW feature computation. Since *bias3* suffers from our suboptimal manual selection of scale parameters for the across-scale differences from [3], it performs worse than the other methods. On the other hand using *bias4* gives best results. Importantly note that a combination of all sampling strategies greatly improves performance.

3.2.2. Mutual Information Analysis

Biased random sampling increases mutual information (MI) and kernel target alignment [34] (KTA) between kernels and labels compared to dense sampling on average over the concepts. This finding is remarkable because these methods use different setups to measure informativeness with respect to labels: Kernel target alignment is based on similarities from Hilbert space

geometry whereas the mutual information criterion relies on co-occurrence probabilities between kernel values and labels. The observed increase in MI and KTA may serve as a theory-driven quantitative validation for our qualitative argument that gradient based local features like SIFT should be sampled over gradient rich regions.

The mutual information $I(K, Y)$ between the kernel K and the labels Y for a concept class has been computed by discretizing the kernel values K_{ij} into F Intervals V_f bounded by equidistant quantiles. Using quantiles ensures that each interval contains the same number of kernel values. We studied several values for $F = 2^6, \dots, 2^{15}$ but the comparison of mutual information between kernels yielded the same qualitative results.

$$I(K, Y) = \sum_{f=1}^F \sum_{YY^\top = +1, -1} p(K_{ij} \in V_f, YY^\top) \log \left(\frac{p(K_{ij} \in V_f, YY^\top)}{p(K_{ij} \in V_f)p(YY^\top)} \right) \quad (6)$$

For kernel target alignment we center the kernel and the labels and computed the quantity

$$A(K, Y) = \frac{\langle K, YY^\top \rangle_F}{\|K\|_F \|YY^\top\|_F} \quad (7)$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius norm. SVMs are invariant to centering due to their dual constraint $\sum_i \alpha_i y_i = 0$. It was argued in [35] that centering is required in order to correctly reflect the test errors from SVMs via kernel alignment.

The centered kernel which achieves a perfect separation of two classes is proportional to $\tilde{\mathbf{y}}\tilde{\mathbf{y}}^\top$, where

$$\tilde{\mathbf{y}} = (\tilde{y}_i), \quad \tilde{y}_i := \begin{cases} \frac{1}{n_+} & y_i = +1 \\ -\frac{1}{n_-} & y_i = -1 \end{cases} \quad (8)$$

and n_+ and n_- are the sizes of the positive and negative classes.

Table 4 shows the mutual information averaged over all concepts and the average kernel target alignment for dense sampling and our three biasedly sampled kernels *bias 1*, *bias 2* and *bias 4*. The absolute values of the mutual information score are small which underlines the difficulty of the task in average but the relative increase for *bias 4* amounts to 9%, for *bias 2* to 10% . This proves our claim about increased mutual information and kernel target alignment under biased random sampling. The mutual information is increased for *bias 4* in 62 concepts and for *bias 2* in 69 concepts out of 99. Apart from considering the average over all classes we can check single concept classes with particularly conspicuous differences under MI and KTA. When considering the classes where dense sampling yields better MI and KTA than biased sampling, the same concepts fill the top four ranks for both methods: *NoBlur*, *PartlyBlurred*, *Sky* and *Clouds*. This is semantically consistent: *Sky* and *Clouds* are lowly textured concepts. Capturing them requires to look at weaker gradients whereas our biased sampling prefers stronger gradients (see the sampling in Figure 3 for an example). Similarly blur is per definition the absence of strong gradients. Thus is is difficult to classify various amounts blur with gradient based sampling methods alone. This shows also the limitations of our biased sampling methods. Another leading concept is *Underexposed* which also shows low gradients because underexposed scenes are overall dark.

On the other end of differences, when considering the classes where biased sampling yields better MI and KTA than dense sampling, eight classes are in the top ten rank for both methods: *Male*, *ParkGarden*, *NoPersons*,

kernel	dense	<i>bias 1</i>	<i>bias 4</i>	<i>bias 2</i>	average
MI ($\times 10^4$)	7.183	7.637	7.838	7.902	7.922
KTA ($\times 10^3$)	23.08	23.00	23.86	23.37	24.19

Table 4: Mutual information and kernel target alignment for average kernels of various sampling types and the average kernel from all types, averaged over all concepts. Same kernels used as in Table 3. Results were computed on training data.

SinglePerson, Portrait, Trees, Plants, LandscapeNature. All are related to vegetation or depictions of humans. For the former we expect texture-rich scenes. For the latter we assume that many person shots have a high contrast to the background and well visible facial features. Both can be captured well using gradient-based biased random sampling. Remarkably, this is semantically consistent as well.

In conclusion it is apparent how differences between dense and biased sampling for single concepts in mutual information and kernel target alignment correspond to our own intuition about the concepts and are consistent between both measures. This demonstrates the sanity of usage of these measures for comparison of kernels.

This analysis has one drawback however: the increase in kernel target alignment and mutual information is not reflected by the SVM results in Table 3. Our hypothesis is that the overfitting property of SVMs may limit the advantage of biasedly sampled kernels with respect to mutual information and kernel target alignment. The key difference between kernel target alignment and the AP score over an SVM output function is that the latter

uses an output computed as a kernel weighted by *learned* support vectors. However it is known that the SVMs overfit strongly on training data relative to testing splits in a cross-validation setup. For many classes we had AP scores on training data of 100 implying perfect separation. Such overfitting may imply a suboptimal selection of support vectors for densely and biasedly sampled kernels which can reduce the advantage of the biasedly sampled kernels compared to the densely sampled kernel. The overfitting may be stronger for the biasedly sampled kernels because we had used less local features per image. This view can be qualitatively supported by the interpretation of the BoW feature as an expectation: the densely sampled kernel used more local features per image, thus the expectation defining a BoW feature has a lower variance and learning might be more stable. This variance reduction effect can be observed experimentally in Table 3: with the exception of the non-gradient based sampling *bias 3* the densely sampled kernel had a lower variance in AP scores over the cross-validation batches than the biasedly sampled kernels.

Despite that averaging the biasedly sampled kernels with the densely sampled kernel improved AP scores significantly as seen in Table 3 and our challenge results.

4. Output Kernel Multi-Task Learning

Utilizing relations between concepts is crucial for human image understanding as integration of information from other classes reduces the uncertainty and ambiguity about the presence of the target class. Incorporating the information about other concepts as an additional cue into the ranking

process is the main goal of our Output Kernel Multi-Task Learning method.

4.1. Output Kernel MTL Method

A general approach to utilize information from other classes is to use existing trained classifiers and combine their outputs using generic methods. Boosting or Multiple Kernel Learning (MKL) [36] would be a natural candidate. The idea of incorporating information from the classifier outputs of other tasks via MKL has been termed *Output Kernel Multi-Task Learning* and recently published [37]. Note that the Output Kernel MTL idea is, in philosophy, close to Lampert and Blaschko [38], however, their procedure cannot be used for image categorization where there are no bounding boxes for most of the concepts.

In our method we tackle the MTL problem by constructing a classifier which incorporates information from the other categories via MKL for each class separately. Our procedure consists of two steps. It is shown in Figure 4. In the first step, shown as 1 in Figure 4, we construct a kernel (called *output kernel*) from the SVM outputs of each task, whereas in the second step denoted by 2 in Figure 4 we combine the information contained in the visual features and the information contained in the output kernel by applying non-sparse MKL. In more formal terms, we compute the output kernels by measuring similarities between the SVM outputs by using the exponential kernel

$$K_c(\mathbf{x}_i, \mathbf{x}_j) = \exp(-[s_c(\mathbf{x}_i) - s_c(\mathbf{x}_j)]^2), \quad (9)$$

where $s_c(\mathbf{x}_i)$ is the score of SVM for the c -th category for the i -th example \mathbf{x}_i . The output kernels from other tasks are combined with the

feature kernels by MKL which learns the optimal weights $\{\beta_j\}_{j=1}^m$ of the combined kernel $\sum_{j=1}^m \beta_j K_j(\mathbf{x}, \bar{\mathbf{x}})$ and the parameters of support vector machine (SVM) simultaneously by employing a generalized p -norm constraint $\|\beta\|_p = \left(\sum_j \beta_j^p\right)^{1/p} \leq 1$. The kernelized form of non-sparse MKL is given as

$$\begin{aligned} \min_{\beta} \max_{\alpha} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,l=1}^n \alpha_i \alpha_l y_i y_l \sum_{j=1}^m \beta_j k_j(x_i, x_l) & (10) \\ \text{s.t.} \quad & \forall_{i=1}^n : 0 \leq \alpha_i \leq C; \quad \sum_{i=1}^n y_i \alpha_i = 0; \\ & \forall_{j=1}^m : \beta_j \geq 0; \quad \|\beta\|_p \leq 1. \end{aligned}$$

For details on the solution of this optimization problem and its primal form we refer to [36] and the SHOGUN package [39]. The parameter C in equation 10 controls the regularization for SVM parameters α_i . The parameter p in the ℓ_p -Norm in equation 10 controls the regularization for the kernel weights. It expresses the degree of trust in the differences in utility between kernels. The boundary case $p = 1$ results in the known sparse MKL which selects only a few kernels and suppresses most others. Choosing $p = 1$ implies the assumption that utility differences between kernels are large and many kernels should be suppressed by giving them zero or small weights. The opposite boundary case $p \rightarrow \infty$ leads to the summed average kernel SVM which implies that utility differences between kernels are small, thus omitting many kernels would drop information and most kernels should be kept in the mixture with approximately equal weights. Note that the amount of the information which is being transferred between classes is controlled by the weights β and does not need to be set a priori and the interactions between tasks do not need to be symmetric.

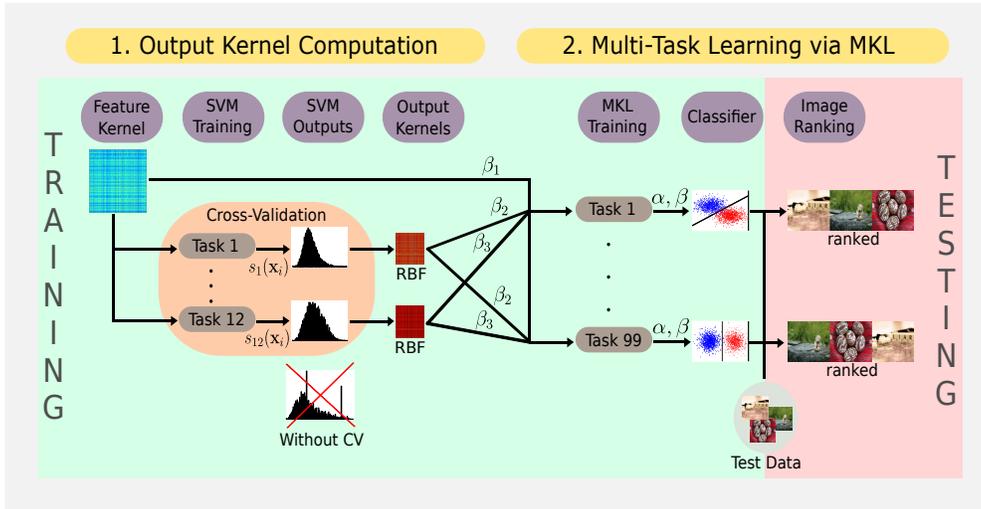


Figure 4: Flow diagram of output kernel MTL method.

As seen in Figure 4 we used in our experiments for computation of output kernels the SVM outputs $s_c(\mathbf{x}_i)$ obtained from cross-validation because SVMs overfit strongly on training data up to the point of perfect predictions which distorts outputs on training data. The Gaussian kernels have been prepared in the same way as described in Section 2. In order to save computation time we did not compute output kernels from all 99 classes, but picked 12 concepts under the constraints to use general concepts and to have rather high mAP values under cross-validation (*Animals, Vehicle, No_Persons, Outdoor, Indoor, Building_Sights, Landscape_Nature, Single_Person, Sky, Water, Sea, Trees*). The classifier outputs have been computed from the average kernel from all kernels in Table 1 which has been also used as the kernel in the *TUBFI 1* submission to the imageCLEF challenge.

We used the 12 output kernels together with the one average kernel from all kernels in Table 1 as inputs for the non-sparse MKL algorithm resulting

in 13 kernels. Please note the line from feature kernel to MKL Training in Figure 4 symbolizing the fact that the visual feature kernel which was computed in an unsupervised manner is part of the MKL kernel set for each concept class. We relied on the MKL implementation from the shogun toolbox [39]. We applied due to an expected lack of time a MKL heuristic. We used one run of MKL using kernel weight regularization parameter $p = 1.125$ and SVM regularization constant $C = 0.75$. The obtained kernel weights were used to compute for each class one weighted sum kernel. This kernel served as input for ordinary SVMs which were trained using a large range of regularization constants $\{10^{-1}, 10^{-1/2}, 1, 10^{+1/2}, 10^{+1}, 0.75, 2\}$.

4.2. Insights from Output Kernel MTL

Here we show the results of our Output Kernel MTL experiments and analyse the reasons of the performance gain. Furthermore we show that our method benefits from using asymmetric relations and visualize the most prominent relations which turn out to be semantically meaningful.

4.2.1. Learning and Asymmetry as Key Features

There are two key features in the Output kernel MTL method – asymmetry and learning of between concept transfer strength. We argue that both properties are essential for a practically useful multi-task method in visual concept ranking when the number of tasks is medium to large. Our argument is rarely seen in the classical application fields of multi-task methods because they are based on different assumptions. Classical application fields of multi-task methods deal typically with data coming with a small number of tasks and small sample sizes where larger sample sizes are often unavailable

or very expensive to obtain e.g. in chemistry or genetics. However in many problems in computer vision data samples and concepts are easier to obtain in larger quantities.

The asymmetry point is obvious for two reasons. We have no desire for a symmetric method. Firstly, given a large number of concepts it is not desirable by computational time considerations to allow every concept to be used as source for information transfer. This implies to learn a number of weights between source and target concepts which is quadratic in the number of concepts. Secondly, when a large set of concepts contains highly varying members, some concepts will have bad recognition rates and low ranking scores. The bad recognition rate of such concepts either will likely spoil the recognition rates of other concepts or the learning method will be robust enough to suppress them. However, bad recognition rates can be identified already during training of single-task base classifiers, so that corresponding concepts can be omitted prior to using them in more complex multi-task learning approaches.

The learning point is less obvious but can be demonstrated on the ImageCLEF dataset. When having 12 source and 87 target concepts – how can we set the 13 times 87 kernel weights? Does class-wise learning matter at all? To see this we tried two alternative ways to set the kernel weights as shown in Table 5. The first way is simply taking the weight for output kernels to be 10% of the weight for the visual kernel. Such a choice expresses the prior assumption that the visual feature kernel is the most informative. The second way uses our MKL weights to obtain the right scale for a kernel mixture: we compute for each source concept class the median weight over all remaining

weights	10% of visual kernel	median of MKL weights	MKL $p = 1.2$
mAP	35.25 ± 7.38	36.16 ± 7.30	36.36 ± 7.35

Table 5: mAP scores for various methods to set weights for output kernels. Results were computed on Training data via 12-fold crossvalidation. The 12 source concept classes are omitted from the mean.

87 target classes and average these medians to assign each output kernel the same weight. Of course, this uses already the results from the MKL learning of kernel weights. Table 5 shows that still the MKL result which learns for each class a separate weight is best. This underlines the variability in the semantic concepts. The MKL weights outperform the second best choice in 62 out of the 87 concepts. The mean average precision is low because we omitted the 12 source classes from the mean computation which have above average mAP scores by crossvalidation on training data.

4.2.2. Ranking Performance Evaluation

We have used the Output Kernel MTL only in submission *TUBFI 4* shown in Table 2. For each task we performed cross-validation on the training data in order to decide whether using Output Kernel MTL is advantageous over the standard approach. By this 42 classes were selected. The first conclusion from Table 2 is that the output MKL gives the biggest gain within all our purely visual submissions. It does not overfit on test data in the sense that relative comparison of performance to other methods has the same sign during the final evaluation test data and from cross-validation on training data. We can see from Figure 5 that using the Output Kernel MTL improves

test performance on most classes over the original visual kernel. Note that the visual kernel is the average from all kernels in Table 1. It has been used as the visual feature kernel in the output kernel MKL and for the *TUBFI 1* submission. Thus transfer of information between classes seems to give additional information and improve ranking performance.

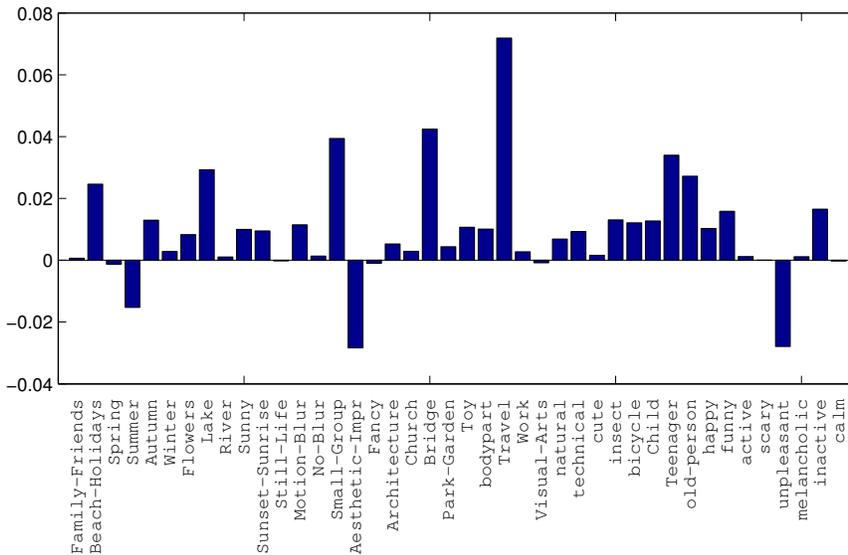


Figure 5: Differences in average precision between output kernel MKL results and the visual baseline kernel which has also been used in output kernel MTL as one kernel. Shown on 42 classes where it has been selected via cross-validation on training data against the visual kernel baseline. Results shown on testing data from the ImageCLEF2011 challenge submission.

4.2.3. Semantic Meaning of Kernel Weights

Another conclusion from our experiments is that the learned semantic relations are intuitively accessible. To see this consider Figure 6 which shows

the fifteen strongest relations between classes as learned by the output kernel MTL in the form of kernel weights. Note firstly that all these relations are asymmetric which could not have been learned using kernel-based MTL methods (e.g. Evgeniou et al. [20]). Each of these relations makes semantic sense. This is a sanity-check for the output kernel MTL and the used BoW features simultaneously. The semantic consistency is in line with the analysis in [37] where we also analysed the relations learned from the 20 PASCAL VOC classes.

We conjecture from Figure 6 that most of the improvement by output kernel MTL can be explained by transfer of *affirmative information* as the *Water* or *Sky* concepts and *eliminative information* depending on the target concept as observed with *No_Persons* source concepts, however, more complex relationships like visual similarity between concepts (or backgrounds) and higher order correlations may also play a role (see [37]).

4.2.4. Mutual Information Analysis

Here we apply the mutual information and kernel target alignment measure in the same manner as done in Section 3.2.2 to assess the information content in our kernels relative to the labels. We compare the visual baseline kernel (see Section 4.1) against the kernel mixtures obtained by the output kernel MTL for each visual concept.

We observe from Table 6 a clear increase in mutual information and kernel target alignment when using the output kernel MTL. The semantic sanity of the underlying relations has been already validated in the Section 4.2.3.

Figure 7 shows the relative differences in mutual information between the baseline kernel and weighted kernel from output kernel MTL for all classes.

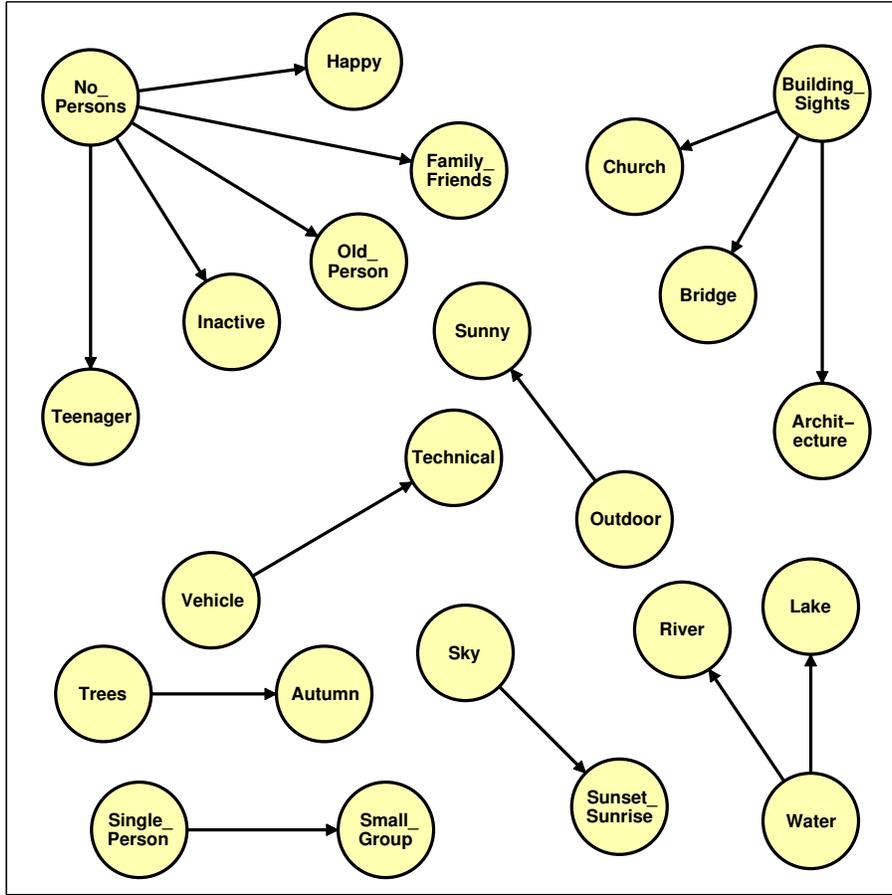


Figure 6: 15 most prominent semantic concept relations obtained using output kernel MKL.

We see an improvement for many concepts. Like in Section 3.2.2, we observe a divergence between the SVM results in Figure 5 and the mutual information, for which we had formulated a conjecture about its cause in Section 3.2.2.

Finally, we would like to show some images which benefit most from the Output Kernel MTL. In Figure 8 we show the images with the largest improvement in rank for the task *Family_Friends* (up left) with old rank:

kernel	baseline	output kernel MTL
MI ($\times 10^4$)	8.42	10.79
KTA ($\times 10^3$)	19.99	25.74

Table 6: Mutual information and kernel target alignment for the baseline kernel and weighted kernel from output kernel MTL averaged over all concepts. Source classes excluded from the average. Scores were computed on training data.

443 and new rank: 236, *Church* (up right) with old rank: 390 and new rank: 236, *Small_Group* (down left) with old rank: 551 and new rank: 346 and *Lake* (down right) with old rank: 324 and new rank: 57. We clearly see that these images benefit from output kernel MTL. Note that three of them are subjectively rather difficult, the church could be a museum as well, the family scene is blurred and the small group is actually a drawing of persons. As shown in Figure 5, in the first case the *No_Persons* classifier helps to identify that there are persons in the image, therefore the *Family_Friends* task becomes easier. In the second case, the fact that there is a building in the image facilitates the ranking for the *Church* concept. According to Figure 6 the *Small_Group* task uses the information of the *Single_Person* classifier and the *Lake* utilizes information from the *Water* classifier. These transfers of information intuitively explain the improvement in ranking in Figure 8.

5. Conclusions

We systematically investigated the impact of two most novel modifications in our successful submissions to the ImageCLEF PhotoAnnotation



Figure 8: Images with largest improvement in rank for certain concepts when using Output Kernel MTL. Upper left: *Family_Friends*. Upper right: *Church*. Lower Left: *Small_Group*. Lower right: *Lake*. Results were computed on training data.

Challenge performed on a part of the MIRFLICKR-1M dataset, namely biased random sampling and output kernel multi-task learning. Biased random sampling based on gradients improves mutual information and kernel target alignment between kernels and labels compared to dense sampling. The differences in mutual information and kernel target alignment between dense sampling and biased sampling across concepts correspond well to the semantics of the pre-defined concepts. Biased random sampling enhances ranking performance. Output kernel multi-task learning yielded the best purely visual submission in the ImageCLEF PhotoAnnotation Challenge.

This method works reliably out of sample and it permits to perform asymmetric information transfer between tasks. It learns the relative importance of source tasks for each target task and scales to training sets of several thousand images. As it is based on multiple kernel learning it is in principle scalable up to datasets with several hundred thousands of training samples as shown in [40] with nonlinear kernels being dealt with e.g. [41]. The learned contributions of source tasks to target tasks are shown to be semantically consistent. The apparent gap between raw mutual information and average precision scores obtained from SVMs may hold interesting future insights on dependencies between representations and learning.

Future work will study a holistic learning procedure where representations and transfer between tasks are part of an overall integrated optimization as well as the impact of both ideas to more complex semantic ranking measures like the ones in [42].

Acknowledgements We like to express our gratitude to Mark Huiskes and Michael S. Lew of the Leiden Institute of Advanced Computer Science <http://www.liacs.nl/> for contributing the dataset, Stefanie Nowak of the ImageCLEF Photo Challenge Organizers from the IDMT in Illmenau, Germany, for definition of the semantic concepts and professional organization of the challenge. Furthermore we like to thank Marius Kloft, Shinichi Nakajima and Volker Tresp. This work was supported by the Federal Ministry of Economics and Technology of Germany (BMWi) under the project THESEUS (01MQ07018) <http://www.theseus-programm.de/>.

About the author—ALEXANDER BINDER obtained a master degree at the Department of Mathematics, Humboldt University Berlin. Since 2007, he has been working for the THESEUS project on semantic image retrieval at Fraunhofer FIRST where he was the principal contributor to top five ranked submissions at ImageCLEF and Pascal VOC challenges. In 2010, he moved to the Machine Learning Group at the TU Berlin and enrolled in their PhD program. His research interests include computer vision, medical applications, machine learning and efficient heuristics.

About the author—WOJCIECH SAMEK obtained a master degree at the Department of Computer Science, Humboldt University Berlin. As a student he worked in the RoboCup project and was visiting scholar at University of Edinburgh and the Intelligent Robotics Group at NASA Ames in Mountain View, CA. He is now a PhD candidate at Berlin Institute of Technology and was awarded a scholarship by the Bernstein Center for Computational Neuroscience. He is also working for the THESEUS Project at Fraunhofer Institute FIRST. His research interests include machine learning, computer vision, robotics, biomedical engineering and neuroscience.

About the author—MOTOAKI KAWANABE obtained a master degree at the Department of Mathematical Engineering, University of Tokyo, Japan. He studied mathematical statistics and received PhD from the same Department in 1995 where he worked as an assistant professor afterwards. He joined the Fraunhofer Institute FIRST in 2000 as a senior researcher. Until fall 2011 he lead the group for the THESEUS project on image annotation and retrieval there. He stayed at Nara Institute of Science and Technology, Kyoto University and RIKEN in Japan in 2007. Now he is with ATR research in Kyoto, Japan. His research interests include computer vision, biomedical data analysis, statistical signal processing and machine learning.

About the author—KLAUS-ROBERT MÜLLER is full Professor for Computer Science at TU Berlin since 2006; at the same time he is directing the Bernstein Focus on Neuro-Technology Berlin. 1999-2006 he was a Professor for Computer Science at Univer-

sity of Potsdam. In 1999, he was awarded the Olympus Prize by the German Pattern Recognition Society, DAGM and in 2006 he received the SEL Alcatel Communication Award. His research interests are intelligent data analysis, machine learning, statistical signal processing and statistical learning theory with the application to computational chemistry, finance, neuroscience, and genomic data. One of his main scientific interests is non-invasive EEG-based Brain Computer Interfacing.

References

- [1] S. Nowak, K. Nagel, J. Liebetrau, The CLEF 2011 photo annotation and concept-based retrieval tasks, in: V. Petras, P. Forner, P. D. Clough (Eds.), CLEF (Notebook Papers/Labs/Workshop), 2011.
- [2] B. T. Mark J. Huiskes, M. S. Lew, New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative, in: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval, ACM, New York, NY, USA, 2010, pp. 527–536.
- [3] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (11) (1998) 1254–1259.
- [4] E. Vazquez, T. Gevers, M. Lucassen, J. van de Weijer, R. Baldrich, Saliency of color image derivatives: A comparison between computational models and human perception, *Journal of the Optical Society of America A* 27 (3) (2010) 613–621.
- [5] T. Liu, J. Sun, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: CVPR, IEEE Computer Society, 2007.
- [6] T. Kadir, M. Brady, Saliency, scale and image description, *International Journal of Computer Vision* 45 (2001) 83–105. doi:10.1023/A:1012460413855.
- [7] N. Rao, J. Harrison, T. Karrels, R. Nowak, T. Rogers, Using machines to improve human saliency detection, in: Signals, Systems and Computers (ASILOMAR), 2010

- Conference Record of the Forty Fourth Asilomar Conference on, 2010, pp. 80–84.
doi:10.1109/ACSSC.2010.5757471.
- [8] F. Moosmann, D. Larlus, F. Jurie, Learning saliency maps for object categorization, in: ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision, Springer, 2006.
URL <http://lear.inrialpes.fr/pubs/2006/MLJ06>
- [9] L. Yang, N. Zheng, J. Yang, M. Chen, H. Chen, A biased sampling strategy for object categorization, in: ICCV, IEEE, 2009, pp. 1141–1148.
- [10] A. Torralba, K. P. Murphy, W. T. Freeman, Sharing visual features for multiclass and multiview object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (5) (2007) 854–869. doi:10.1109/TPAMI.2007.1055.
- [11] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: CVPR, 2009.
- [12] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classemes, in: European Conference on Computer Vision (ECCV), 2010, pp. 776–789.
- [13] X.-T. Yuan, S. Yan, Visual classification with multi-task joint sparse representation, in: IEEE Conf. on Comp. Vision & Pat. Rec., 2010, pp. 3493–3500.
- [14] Y. Wang, G. Mori, A discriminative latent model of object classes and attributes, in: European Conference on Computer Vision, 2010.
- [15] M. Fink, Object classification from a single example utilizing class relevance metrics, in: In Advances in Neural Information Processing Systems (NIPS, MIT Press, 2004, pp. 449–456.
- [16] E. Bart, S. Ullman, Single-example learning of novel classes using representation by similarity, in: In British Machine Vision Conference (BMVC, 2005.
- [17] L. Fei-Fei, Knowledge transfer in learning to recognize visual object classes., ICDL.

- [18] T. Tommasi, B. Caputo, The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories, in: BMVC, 2009.
- [19] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, USA, 2011, pp. 1641–1648.
- [20] T. Evgeniou, C. A. Micchelli, M. Pontil, Learning multiple tasks with kernel methods, *Journal of Machine Learning Research* 6 (2005) 615–637.
- [21] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research* 6 (2005) 1453–1484.
- [22] Z. Song, Q. Chen, Z. Huang, Y. Hua, S. Yan, Contextualizing object detection and classification, in: CVPR, IEEE, 2011, pp. 1585–1592.
- [23] K. Ming, A. Chai, C. K. I. Williams, S. Klanke, S. Vijayakumar, Multi-task gaussian process learning of robot inverse dynamics, in: *Neural Inf. Proc. Sys.*, 2008.
- [24] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories., in: *Workshop on Generative-Model Based Vision*, 2004.
- [25] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, *IEEE Transactions on Neural Networks* 12 (2) (2001) 181–201.
- [26] C. Dance, J. Willamowski, L. Fan, C. Bray, G. Csurka, Visual categorization with bags of keypoints, in: *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.
URL http://www.xrce.xerox.com/Publications/Attachments/2004-010/2004_010.pdf

- [27] M. Marszalek, C. Schmid, Learning representations for visual object class recognition.
URL <http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/workshop/marszalek.pdf>
- [28] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating color descriptors for object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (9) (2010) 1582–1596.
- [29] J. van Gemert, J.-M. Geusebroek, C. J. Veenman, A. W. M. Smeulders, Kernel codebooks for scene categorization, in: D. A. Forsyth, P. H. S. Torr, A. Zisserman (Eds.), *ECCV* (3), Vol. 5304 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 696–709.
- [30] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60 (2) (2004) 91–110.
- [31] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), *Comput. Vis. Image Underst.* 110 (2008) 346–359. doi:10.1016/j.cviu.2007.09.014.
- [32] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: A. Leonardis, H. Bischof, A. Pinz (Eds.), *ECCV* (4), Vol. 3954 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 490–503.
- [33] K. E. A. van de Sande, T. Gevers, The University of Amsterdam’s concept detection system at ImageCLEF 2010, in: M. Braschler, D. Harman, E. Pianta (Eds.), *CLEF (Notebook Papers/LABs/Workshops)*, 2010.
- [34] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. S. Kandola, On kernel-target alignment, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), *NIPS*, MIT Press, 2001, pp. 367–373.
- [35] C. Cortes, M. Mohri, A. Rostamizadeh, Two-stage learning kernel algorithms, in: J. Fürnkranz, T. Joachims (Eds.), *ICML*, Omnipress, 2010, pp. 239–246.

- [36] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Lp-norm multiple kernel learning, *Journal of Machine Learning Research* 12 (2011) 953–997.
- [37] W. Samek, A. Binder, M. Kawanabe, Multi-task learning via non-sparse multiple kernel learning, in: P. Real, D. Díaz-Pernil, H. Molina-Abril, A. Berciano, W. G. Kropatsch (Eds.), *CAIP* (1), Vol. 6854 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 335–342.
- [38] C. Lampert, M. Blaschko, A multiple kernel learning approach to joint multi-class object detection, in: *DAGM*, 2008, pp. 31–40.
- [39] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. D. Bona, A. Binder, C. Gehl, V. Franc, The shogun machine learning toolbox, *Journal of Machine Learning Research* 11 (2010) 1799–1802.
- [40] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf, Large scale multiple kernel learning, *Journal of Machine Learning Research* 7 (2006) 1531–1565.
- [41] A. Rahimi, B. Recht, Random features for large-scale kernel machines, in: J. C. Platt, D. Koller, Y. Singer, S. T. Roweis (Eds.), *NIPS*, MIT Press, 2007.
- [42] A. Binder, K.-R. Müller, M. Kawanabe, On taxonomies for multi-class image categorization, *International Journal of Computer Vision* (2011) 1–21doi:10.1007/s11263-010-0417-8.