The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task

Alexander Binder[1], Wojciech Samek[1,2], Marius Kloft[1], Christina Müller[1], Klaus-Robert Müller[1], and Motoaki Kawanabe[2,1]

[1] Machine Learning Group, Berlin Institute of Technology (TU Berlin), Franklinstr. 28/29, 10587, Berlin, Germany, www.ml.tu-berlin.de
alexander.binder@tu-berlin.de, wojciech.samek.tu-berlin.de
[2] Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany
motoaki.kawanabe@first.fraunhofer.de

Abstract. In this paper we present details on the joint submission of TU Berlin and Fraunhofer FIRST to the ImageCLEF 2011 Photo Annotation Task. We sought to experiment with extensions of Bag-of-Words (BoW) models at several levels and to apply several kernel-based learning methods recently developed in our group. For classifier training we used non-sparse multiple kernel learning (MKL) and an efficient multi-task learning (MTL) heuristic based on MKL over kernels from classifier outputs. For the multi-modal fusion we used a smoothing method on tag-based features inspired by Bag-of-Words soft mappings and Markov random walks. We submitted one multi-modal run extended by the user tags and four purely visual runs based on Bag-of-Words models. Our best visual result which used the MTL method was ranked first according to mean average precision (MAP) within the purely visual submissions. Our multi-modal submission achieved the first rank by MAP among the multi-modal submissions and the best MAP among all submissions. Submissions by other groups such as BPACAD, CAEN, UvA-ISIS, LIRIS were ranked closely. For more details we refer to our publication accepted to the Computer Vision and Image Understanding journal.

Keywords: ImageCLEF, Photo Annotation, Image Classification, Bag-of-Words, Multi-Task Learning, Multiple Kernel Learning, THESEUS

1 Introduction

Our goals were to experiment with extensions of Bag-of-Words (BoW) models at several levels and to combine them with several kernel-based learning methods recently developed in our group while working within the THESEUS project. For this purpose we generated a submission to the annotation task of the ImageCLEF2011 Photo Annotation Challenge [14]. This task required the annotation of 10000 images in the provided test corpus according to the 99 pre-defined categories. Note that this year's ImageCLEF Photo-based task provides additionally another challenging competition [14], a concept-based retrieval task. In the following we will focus on the firstly mentioned annotation task over the 10000 images. The ImageCLEF photo corpus is challenging

Table 1: BoW Feature Sets. See text for explanation.

Sampling Type	Local Feature	Color Channels	BoW Mapping	No. of Features
grid	Color Quantiles	RGB, Opp,Gr	Rank	9
grid	SIFT	RGB, Opp,Gr, N-Opp	0-1	12
grid	SIFT	RGB, Opp,Gr, N-Opp	Rank	12
bias1	SIFT	RGB, Opp,Gr	Rank	9
bias2	SIFT	RGB, Opp,Gr, N-Opp,N-RGB	Rank	15
bias3	SIFT	RGB, Opp	Rank	6
bias4	SIFT	RGB, Opp,Gr	Rank	9

due to its heterogeneity of classes. It contains classes based on concrete tangible objects such as female, cat and vehicle as well as more abstractly defined classes such as technical, boring or Esthetic_Impression. As a result our visual submission and our multi-modal submission achieved both first ranks by MAP measure among the purely visual and multi-modal submissions, respectively. We will describe our methods in a concise manner here.

2 Bag-of-Words Features

All our submissions were based on discriminatively trained classifiers over kernels using BoW features. The BoW feature pipeline can be decomposed into the following steps: generating sampling regions, computing local features, mapping local features onto visual words. The coarse layout of our approach is influenced by the works of the Xerox group on Bag-of-Words in Computer Vision [3], the challenge submissions by INRIA groups [11] and the works on color descriptors by the University of Amsterdam [17]. For that reason we computed for each set of parameters three BoW features based on regular spatial tilings $1 \times 1, 2 \times 2, 3 \times 1$ (vertical $\times$ horizontal). Preliminary experiments with an additional spatial tiling 3×3 showed merely minor performance gains. Furthermore we used vectors of quantile estimators along the established SIFT feature [10] as local feature. Table 1 shows the computed BoW features. Information about the sampling method is given in Section 2.1. We used color channel combinations red-green-blue (RGB), grey (Gr), grey-opponentcolor1-opponentcolor2 (Opp in Table 1) and a grey-value normalized version of the last combination (N-Opp in Table 1). The total number of kernels is large however their computation is a fairly automatized task which requires little human intervention.

In this years submission, we incorporated the following new extensions described in Sections 2.1 and 2.2 into our BoW modeling.

2.1 Extensions on Sampling level

In addition to BoW features created from known grid sampling we tried biased random sampling [21]. In contrast to [21] we resorted to probability maps computed from edge detectors. Such sampling approaches offer two potential advantages over Harris Laplace detectors: Firstly, we get keypoints located on edges rather than corners. A motivating

example can be seen in Figure 1 – 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 circular structure.

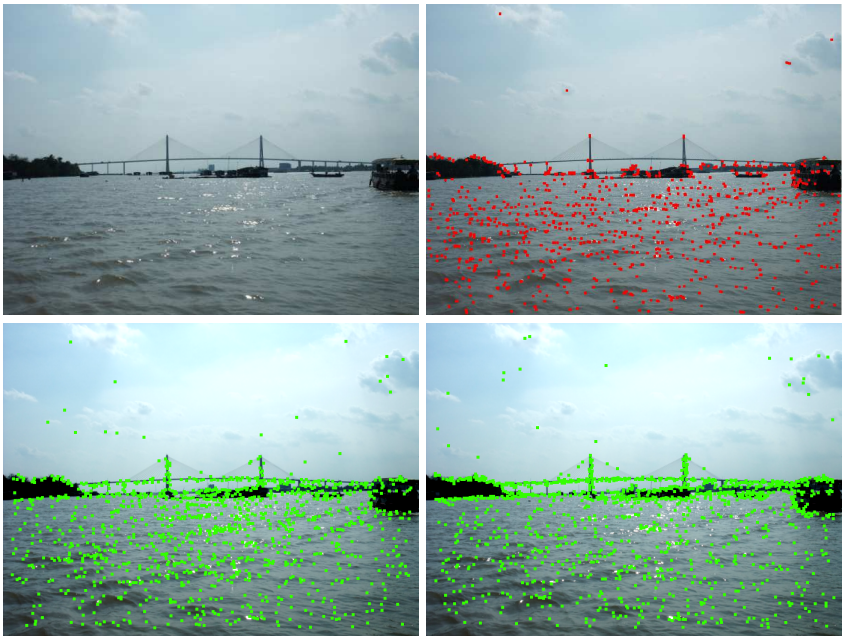

Fig. 1: Upper Left: The essential structures of the bridge are lines rather than corners (author's own work). Upper Right: Harris Laplace keypoints. Lower Left: bias1 keypoints. Lower Right: bias4 keypoints with same number of keypoints as detected by Harris Laplace.

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 the number of local features extracted by grid sampling, in our case 6 pixels. This comes from the idea that some images can be more smooth in general. Furthermore [13] showed that too sparse sampling of local features leads to reduced classification performance. The opposite extreme end of this is documented in [18] where quite large improvements using sampling each pixel are reported. As a consequence we can tune the trade-off between computational cost and performance compared to the dense sampling baseline. In practice we chose to extract approximately one half as much local features using biased random sampling. We tried four detectors:

– *bias3* was a simplified version of an attention based detector [7]. However this detector requires to set scale parameters. The highly varying scales of motifs in the images makes it difficult to find a globally optimal set of scales without expensive optimizations. This inspired us to try detectors which depend less 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 blue minus red.
- *bias2* is like bias1 except for dropping the grey channel. 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 scale as the SIFT descriptor. Regions with low entropy are preferred in the probability map used for biased random sampling. This detector is adapted closely to the SIFT feature. The question behind this detector is whether the focus on peaky low entropy histograms constitutes an advantage.

2.2 Extensions on Bag-of-Words Mapping Level

As we used k-means for generating a set of visual words, the usual approach to generate soft BoW mappings [5] which is adapted to radius-based clustering and relies on one global width parameter may become inappropriate when the density of clusters varies strongly in the space of local features. K-means results in clusters of varying size depending on the local density of the local features. To resolve this issue we resorted to rank-based BoW mapping where the vote of a local feature is the 2.4-based power of the negative rank. Be $RK_d(l)$ the rank of the distances between the local feature l and the visual word corresponding to BoW dimension d, sorted in increasing order. Then the BoW mapping m_d for dimension d is defined as:

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

Initially we performed experiments with several alternative soft mappings. Shortly summarized, these experiments revealed that it is necessary to achieve a sufficiently fast decay of soft mapping weights as a function of the distance of a local feature to distant visual words in order to achieve a better performance than simple hard mapping.

Our second attempt after using the mapping from [5] was to introduce a cutoff constant K. Only distances below rank $K+1$ are considered. Be $\mathcal{V}$ a visual vocabulary, and w_d the visual word from it corresponding to BoW feature dimension d, l a local feature. Then the cut-off mapping is given by:

$$m_d(l) = \begin{cases} \frac{\exp(-\sigma_{w_d}\mathrm{dist}(l,w_d))}{\sum_{v\in\mathcal{V}|\mathrm{Rank}(\mathrm{dist}(l,v))\leq K}\exp(-\sigma_v\mathrm{dist}(l,v))} & \text{if } \mathrm{Rank}(\mathrm{dist}(l,w_d)) \leq K \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where the width parameter σ was estimated for each visual word locally as the inverse of quantile estimators of distances to all local features from an image which had w_d as the nearest visual word. Trying out factors of 0.1,1 and 10 on the median of these distances proved effective in practice.

This experiment led to the conclusion that quantiles leading to large values for σ and thus fast decay of weights yielded better performances.

Note that the rank-based voting ensures exponential drop-off per se.

2.3 Kernels

We used χ^2-Kernels. The width was set to be the mean of the inner distances.

2.4 Used Resources

For feature and kernel computations we resorted to a cluster with 40 mostly AMD Opterons 275 Core Units with up to 2.4 GHz which had according to `cpubenchmark.net` a speed rank of 134 in August 2011. The OS was a 32bit which limited usable memory resources during feature computation, in particular during visual word generation to 3 GByte.

3 Heuristically down-scaled non-sparse Multiple Kernel Learning

Due to limited resources on a 64 bit cluster which we employed for classifier training we decided to try out a down-scaled version of MKL based on 25 kernels which are the averages of the 75 kernels over the spatial tilings. Instead of evaluating many pairs of sparsity parameters and regularization constants the idea was to run non-sparse MKL [9] once for each class for merely one sparsity parameter tuned towards low kernel weight regularization ($p = 1.2$) and one choice of the regularization constant tuned towards high SVM regularization ($C = 0.1$). The obtained kernel weights can be used afterwards in SVMs with fixed-weighted kernels and several weaker SVM regularizations and powers applied to the kernel weights simulating higher sparsity. This consumes substantially less memory and allows in practice to use more cores in parallel. For each class one can choose via cross-validation the optimal regularization and power on the initially obtained MKL weights.

4 Output Kernel based MKL/MTL

By considering the set of semantic concepts in the ImageCLEF Photo one can expect weak relations between many of them. Some of them can be established deterministically such as season labels like Spring necessarily require the photo to be an outdoor shot. Others might be present in a statistical sense: photos showing Park_Garden tend to be rather calm instead of active, however the latter is possible. The extent of activity might depend on the dataset. The total number of concepts is however prohibitive for manual modeling of all relations. One principled approach for exploiting such relations is multi-task learning [4, 19] which attempts to transfer information between concepts. Classical Multi-Task Learning (MTL) has two shortcomings: firstly, it often scales poorly with the number of concepts and samples. Secondly, kernel-based MTL leads to symmetric solutions, which implies that poorly recognized concepts can spoil

classification rates of better performing classes. The work in [16] tackles both problems. It formulates a decomposable approximation which can be solved as a set of separate MKL problems. Thus it shares the scalability limits of MKL approaches. Secondly, the formulation as an approximation permits asymmetric information transfer between classes. The approach uses kernels computed from SVM predictions for the information transfer. We picked 12 concepts under the constraints to use general concepts and to have rather high MAP values under cross-validation for the kernels (*animals*, *food*, *no_persons*, *outdoor*, *indoor*, *building_sights*, *landscape_nature*, *single_person*, *sky*, *water*, *sea*, *trees*) and combined them with the average kernel which has been used in the TUBFI 1 submission as inputs for the non-sparse MKL algorithm resulting in 13 kernels. Here we applied as a consequence of lack of computation time the same downscaled MKL strategy as for the BoW kernels alone described in Section 3 with MKL regularization parameter $p = 1.125$ and SVM regularization constant $C = 0.75$, however without applying sparsifying powers on the kernel weights.

5 Smoothing textual Bag-of-Words

In the field of image classification it is known that soft mappings improve the performance of BoW features substantially [5, 20]. The soft mapping relies on a notion of distance between the local features. Similar approaches have been also used for Fisher vectors [12, 15] where the non-sparsity of features does not require a distance. For tag based BoW features one can derive analogously a notion of similarity without resorting to external sources via co-occurrence. We applied for our multi-modal submission TUBFI 3 the method from [8] which uses derived similarity to achieve a soft mapping for textual bags of words. The set of visual words has been selected by choosing the 0.2% most frequent tags as in [6]. Experiments on the cross-validated training set confirmed performance improvements using smoothed tags over unsmoothed tags.

6 Results

For the detailed results of all submissions we refer to the overview given in [14]. A small excerpt can be seen in Table 2.

While optimization of complex measures, particularly hierarchically representable ones is feasible [1] we did not do any optimization for the example-based measures as we were not fully aware of their structure. In particular, each classifier had a different threshold due to a simple linear mapping of minimal and maximal outputs onto boundaries of the required interval $[0, 1]$ which leaves the targeted MAP values unchanged. This across-concept variation in the classifier threshold explains the limited results for the example-based measures.

Considering the targeted MAP score, we can see in Table 2 that the pure textual runs perform worst although one can expect them to be very efficient in terms of time consumption versus ranking performance difference to visual ones. Multi-modal approaches perform best with a considerable margin of 0.06 MAP (16% over TUBFI 1 baseline) over visual ones which indicates that the information in tags and images is fairly non-redundant. The improvement over pure textual runs is substantial. When

considered as an absolute number, an MAP of 44 shows much space for improvements. When looking at AUC values which allow better comparison between concepts, only 31 out of 99 classes had AUC values above 0.9 in our best submission.

Table 2: Results by MAP for the best three submissions.

Submission	Modality	MAP on test data
BPACAD 3	T	**34.56**
IDMT 1	T	32.57
MLKD 1	T	32.56
TUBFI 1	V	38.27
TUBFI 2	V	37.15
TUBFI 4	V	**38.85**
TUBFI 5	V	38.33
CAEN 2	V	38.24
ISIS 3	V	37.52
TUBFI 3	V&T	**44.34**
LIRIS 5	V&T	43.70
BPACAD 5	V&T	43.63

7 Discussion of Submission Outcomes

The first purely visual submission (TUBFI 1 in Table 2) was an average kernel SVM over all sets but the second BoW feature set from Table 1. Its performance was almost identical to the best submission of the CAEN group which, however, used a completely different methodology, namely Fisher-Kernels. For all other submissions we selected for each class separately the best classifier from a set of classifiers by MAP values obtained on 12-fold cross-validation on the training data. The idea was to counter the heterogeneity of concept classes in the ImageCLEF data by a bag of methods. However, this mixture does not allow to judge the impact of the separate methods precisely. Table 3 shows for each submission applied the number of classes using particular method. Selection was based on cross-validated MAP.

The pool for the second purely visual submission (TUBFI 2 in Table 2) consisted of average kernels computed over several combinations of the sets from Table 1. Hypothesis testing using a Wilcoxon's signed rank test on the cross-validation results showed no improvement in MAP over the first pure visual submission. Nevertheless we submitted it for the sake of scientific curiosity – using average kernel SVMs over varying sets of kernels is a computationally very efficient method. On the test data we observed a drop in performance. Table 3 shows for each submission applied the number of classes using particular method.

The pool for the fourth purely visual submission (TUBFI 5 in Table 2) consisted of the classifiers from the second submission combined with a MKL heuristic (see Section 3). Statistical testing revealed that a small improvement in MAP could be expected.

Indeed, the result on test data was marginally better than the first and the second submission despite it contained some classifiers from the flawed second submission and used a heuristically down-scaled variant of MKL.

The pool for the third and a posteriori best purely visual submission (TUBFI 4 in Table 2) consisted of the classifiers from the second submission, the MKL heuristic and the output kernel MKL/MTL (see Section 4). Statistical testing revealed more classes with significant gains over the baseline (TUBFI 1 in Table 2). In 42 categories the chosen classifier belongs to the output kernel MKL/MTL. The result on test data was the only purely visual run which showed a larger improvement over the baseline TUBFI 1. Therefore we attribute its gains to the influence of the output kernel MKL procedure.

The only multi-modal submission (TUBFI 3 in Table 2) used kernels from the baseline (TUBFI 1 in Table 2) combined with smoothed textual Bag-of-Words (see Section 5).

Table 3: number of classes using a particular method shown for each submission.

submission	TUBFI 1 kernels	other kernel sets	MKL heur.	Output MKL heur.	tag kernels
TUBFI 1	99	0	0	0	0
TUBFI 2	35	64	0	0	0
TUBFI 5	5	27	67	0	0
TUBFI 4	1	16	40	42	0
TUBFI 3	6	0	0	0	93

Acknowledgments We like to thank Shinichi Nakajima, Roger Holst, Dominik Kuehne, Malte Danzmann, Stefanie Nowak, Volker Tresp and Klaus-Robert Müller. This work was supported in part by the Federal Ministry of Economics and Technology of Germany (BMWi) under the project THESEUS (01MQ07018).

References

1. Binder, A., Müller, K.R., Kawanabe, M.: On taxonomies for multi-class image categorization. International Journal of Computer Vision pp. 1–21 (January 2011), http://dx.doi.org/10.1007/s11263-010-0417-8
2. Braschler, M., Harman, D., Pianta, E. (eds.): CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy (2010)
3. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22. Prague, Czech Republic (May 2004)
4. Evgeniou, T., Micchelli, C.A., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research 6, 615–637 (2005)
5. van Gemert, J., Geusebroek, J., Veenman, C., Smeulders, A.: Kernel codebooks for scene categorization. In: ECCV (2008)
6. Guillaumin, M., Verbeek, J., Schmid, C.: Multimodal semi-supervised learning for image classication. In: Proc. of IEEE Int. Conf. on Comp. Vis. & Pat. Rec. (CVPR '10). San Francisco, CA, USA (2010), http://lear.inrialpes.fr/pubs/2010/GVS10

7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20(11), 1254–1259 (1998)
8. Kawanabe, M., Binder, A., Müller, C., Wojcikiewicz, W.: Multi-modal visual concept classification of images via markov random walk over tags. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on. pp. 396 –401 (2011)
9. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: Lp-norm multiple kernel learning. Journal of Machine Learning Research 12, 953–997 (Mar 2011)
10. Lowe, D.: Distinctive image features from scale invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
11. Marszalek, M., Schmid, C.: Learning representations for visual object class recognition, http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/workshop/marszalek.pdf
12. Mensink, T., Csurka, G., Perronnin, F., Sánchez, J., Verbeek, J.J.: Lear and xrce's participation to visual concept detection task - imageclef 2010. In: Braschler et al. [2]
13. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV (4). Lecture Notes in Computer Science, vol. 3954, pp. 490–503. Springer (2006)
14. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. In: CLEF 2011 working notes. The Netherlands (2011)
15. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV (4). Lecture Notes in Computer Science, vol. 6314, pp. 143–156. Springer (2010)
16. Samek, W., Binder, A., Kawanabe, M.: Multi-task learning via non-sparse multiple kernel learning. In: CAIP (2011), accepted
17. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pat. Anal. & Mach. Intel. (2010)
18. van de Sande, K.E.A., Gevers, T.: The university of amsterdam's concept detection system at imageclef 2010. In: Braschler et al. [2]
19. Sheldon, D.: Graphical multi-task learning. http://agbs.kyb.tuegingen.mpg.de/wikis/bg/siso2008/Sheldon.pdf (2008), nIPS workshop on strictured input - structured output
20. Tahir, M., van de Sande, K., Uijlings, J., Yan, F., Li, X., Mikolajczyk, K., Kittler, J., Gevers, T., Smeulders, A.: SurreyUVA SRKDA method. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/tahir.pdf
21. Yang, L., Zheng, N., Yang, J., Chen, M., Chen, H.: A biased sampling strategy for object categorization. In: ICCV. pp. 1141–1148 (2009)