Enhancing Recognition of Visual Concepts with Primitive Color Histograms via Non-sparse Multiple Kernel Learning

Alexander Binder[1,2] and Motoaki Kawanabe[1,2]

[1] Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany
alexander.binder@tu-berlin.de, motoaki.kawanabe@first.fraunhofer.de
[2] TU Berlin, Franklinstr. 28/29, 10587 Berlin, Germany

Abstract. In order to achieve good performance in image annotation tasks, it is necessary to combine information from various image features. In recent competitions on photo annotation, many groups employed the bag-of-words (BoW) representations based on the SIFT descriptors over various color channels. In fact, it has been observed that adding other less informative features to the standard BoW degrades recognition performances. In this contribution, we will show that even primitive color histograms can enhance the standard classifiers in the ImageCLEF 2009 photo annotation task, if the feature weights are tuned optimally by non-sparse multiple kernel learning (MKL) proposed by Kloft et al.. Additionally, we will propose a sorting scheme of image subregions to deal with spatial variability within each visual concept.

1 Introduction

The original publication is available at http://www.springerlink.com in the LNCS series volume 6242 for the ImageCLEF2009 PostProceedings published by the Springer company at
http://www.springerlink.com/content/978-3-642-15750-9/#section=783177\&page=1.

Recent research results show that combining information from various image features is inevitable to achieve good performance in image annotation tasks. With the support vector machine (SVM) [1,2], this is implemented by mixing kernels (similarities between images) constructed from different image descriptors with appropriate weights. For instance, the average kernel with uniform weights or the optimal kernel trained by multiple kernel learning (called ℓ^1-MKL later) have been used so far. Since the sparse ℓ^1-MKL tends to overfit by ignoring quite a few kernels, Kloft et al. [3] proposed the non-sparse MKL with ℓ^p-regularizer ($p \geq 1$), which bridges the average kernel ($p = \infty$) and ℓ^1-MKL. The non-sparse MKL is successfully applied to object classification tasks; it could outperform the two baseline methods by optimizing the tuning parameter $p \geq 1$ through cross validation. In particular, it is useful to combine less informative features such as color histograms with the standard bag of words (BoW) representations [4]. We

will show that by ℓ^p-MKL additional simple features can enhance classification performances of some visual concepts in the ImageCLEF 2009 photo annotation task [5], while with the average kernel they just degrade recognition rates. Since the images are not aligned, we will also propose a sorting scheme of image subregions to deal with the spatial variability, when computing similarities between different images.

2 Features and Kernels Used in our Experiments

Features For the following experiments, we prepared two kinds of image features: one is the BoW representations based on the SIFT descriptors [6] and the other is the pyramid histograms [7] of color intensities (PHoCol). The BoW features were constructed in a standard way. By the code used in [8], the SIFT descriptors were computed on a dense grid of step size six over multiple color channels: red, green, blue, and grey. Then, for both grey and combined red-green-blue channels, 4000 visual words (prototypes) were generated by using k-means clustering with large sets of SIFT descriptors selected randomly from the training images in analogy to [9]. For each image, one of the visual words was assigned to the base SIFT at each grid point and the set of words was summarized in a histogram within each cell of the spatial tilings 1×1, 2×2 and 3×1 [7]. Finally, we obtained 6 BoW features (2 colors $\times$ 3 pyramid levels). On the other hand, the PHoCol features were computed by making histograms of color intensities with 10 bins within each cell of the spatial tiling 4×4 and 8×8 for various color channels: grey, opponent color 1, opponent color 2, normalized red, normalized green, normalized blue. The finer pyramid levels were considered, because the intensity histograms usually contain only little information.

Sorting the color histograms The spatial pyramid representation [7] is very useful, in particular, when annotating aligned images, because we can incorporate spatial relations of visual contents in images properly. However, if we want to use histograms on higher-level pyramid tilings (4×4 and 8×8) as parts of input features for general digital photos of the annotation task, it is necessary to handle large spatial variability within each visual concept. Therefore, we propose to sort the cells within a pyramid tiling according to the slant of the histograms. Mathematically, our sort criterion $sl(h)$ is defined as

$$\widetilde{a}[h]_i = \sum_{k \leq i} h_k, \quad a[h]_i = \frac{\widetilde{a}[h]_i}{\sum_k \widetilde{a}[h]_k}, \quad sl(h) = -\sum_i a[h]_i \ln(a[h]_i). \tag{1}$$

The idea behind the criterion can be explained intuitively. The accumulation process $a[h]$ maps the histogram h with only one peak at the highest intensity bin to the minimum entropy distribution (Fig. 1 left) and that with only one peak at the lowest intensity bin to the maximum entropy distribution (Fig. 1 right). If the original histogram h is flat, the accumulated histogram $a[h]$ becomes a

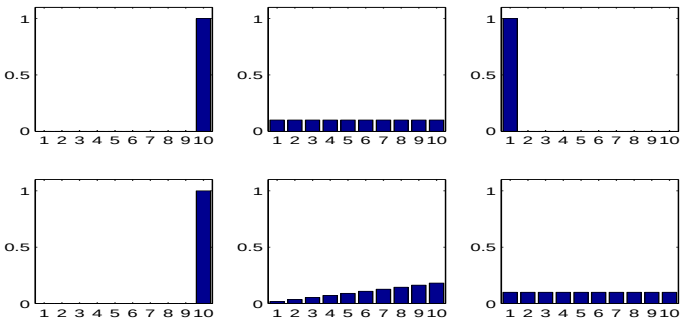

Fig. 1. Explanation of the slant score. Upper: intensity histograms. Lower: corresponding histograms accumulated.

linearly increasing function which has an entropy in between the two extremes (Fig 1 middle).

On the other hand, it is natural to think that all possible permutations π are not equally likely in sorting of the image cells. In many cases, spatial positions of visual concepts can change more horizontally than vertically (e.g. sky, sea). Therefore, we introduced a sort cost in order to punish large changes of the vertical positions of the image cells before and after sorting.

$$sc(\pi) = C \sum_{k} \max(v(\pi(k)) - v(k), 0) \tag{2}$$

Here $v(i)$ denotes the vertical position of the i-th cell within a pyramid tiling and the constant C is chosen such that the sort cost is upper-bounded by one and lies at a similar range compared to the χ^2-distance between the color histograms. The sort cost is used to modify the PHoCol kernels. When comparing images x and y, the squared distance between the sort costs are added to the χ^2-distance between the color histograms.

$$k(x, y) = \exp[-\sigma\{d_{\chi^2}(h_x, h_y) + (sc_x - sc_y)^2\}] \ . \tag{3}$$

In our experiments, we computed the sorted PHoCol features on 4×4 and 8×8 pyramid tilings and constructed kernels with and without the sort cost modification. Although the intensity-based features have a lesser performance as standalone image descriptors even after the sorting modifications, combining them with the standard BoW representations can enhance performances in some of the 53 classification problems in the ImageCLEF09 task with almost no additional computation costs.

Kernels We used the χ^2-kernel except for the cases that the sort cost was incorporated. The kernel width was set to be the mean of the inner χ^2-distances computed over the training data. All kernels were normalized.

3 Experimental Results

We aim at showing an improvement over a gold standard represented by BoW features with average kernel while lacking the ground truth on the test data. Therefore we evaluated all settings using 10-fold cross validation on the ImageCLEF09 photo annotation training data, consisting of 5000 images. This allows to perform statistical testing and to predict generalization errors for selecting better methods/models. The BoW baseline is a reduced version ouf our ImageCLEF submission described in the working notes. The submitted version gave rise to results behind the ISIS and INRIA-LEAR groups by AUC (margins 0.022, 0.006) and by EER also behind CVIUI2R (margins 0.02, 0.005, 0.001). XRCE and CVIUI2R performed better by the hierarchy measure (margins 0.013, 0.014). We report in this section performance comparison between SVMs with the average kernels, the sparse ℓ^1-MKL, and the non-sparse ℓ^p-MKL [3]. In ℓ^p-MKL, the tuning parameter p is selected for each class from the set $\{1.0625, 1.125, 1.25, 1.5, 2\}$ by cross validation scores and the regularization parameter of the SVMs was fixed to one.

We chose the average precision (AP) as the evaluation criterion which is also employed in the Pascal VOC Challenges due to its sensitivity to smaller changes, even when AUC values are already saturated above 0.9. This rank-based measure is invariant against the actual choice of a bias. We did not employ the equal error rate (EER), because it can suffer from unbalanced sizes of the ImageCLEF09 annotations. We remark that several classes have less than 100 positive samples and generally no learning algorithm generalizes well in such cases.

We will pose four questions and present experimental results to answer them in the following.

Does MKL help for combining the bag of words features? Our first question is whether the MKL techniques are useful compared to the average kernel SVMs for combining the default 6 BoW features. The upper panel of Fig. 2 shows the performance differences between ℓ^p-MKL with class-wise optimal p's and SVMs with the average kernels over all 53 categories. The classes are sorted as in the guidelines of the ImageCLEF09 photo annotation task.

In general, we see just minor improvements by applying ℓ^p-MKL in 33 out of 53 classes and for only one class it achieved major gain. Seemingly, the chosen BoW features have on average similar information contents. The cross-validated scores (average AP 0.4435 and average AUC 0.8118) of the baseline method imply that these 6 BoW features contributed mostly to the final results of our group achieved on the testdata of the annotation task.

On the other hand, the lower panel of Fig. 2 indicate that the canonical ℓ^1-MKL is not a good idea in this case. On average over all classes ℓ^1-MKL gives worse results compared to the baseline. We attribute this to the harmful effects of sparsity in noisy image data.

Our observations are quantitatively supported by Wilcoxon signed rank test (the significance level $\alpha = 0.05$) which can tell the significance of the perfor-

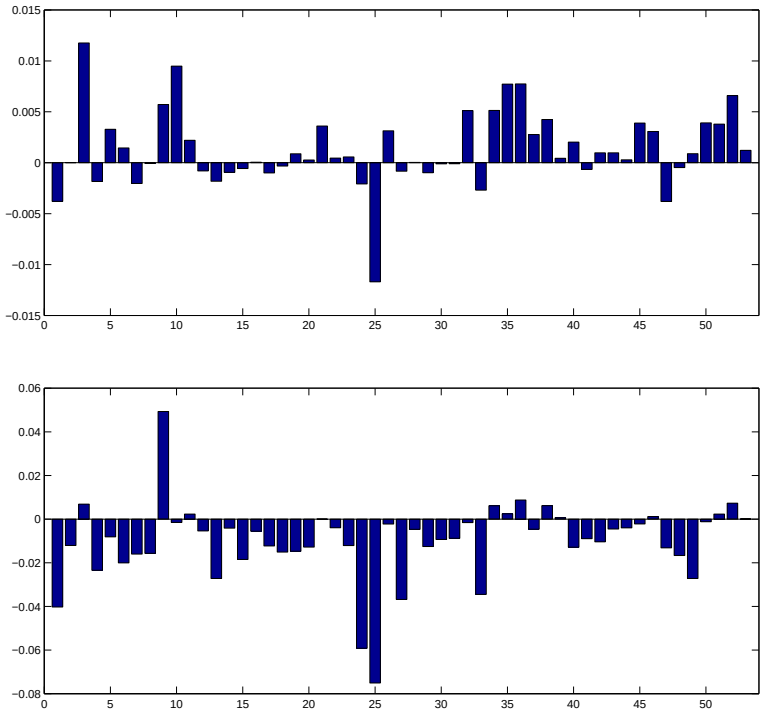

Fig. 2. Class-wise performance differences when combining the 6 BoW features. Upper: ℓ^p-MKL vs the average kernel SVM. Lower: ℓ^1-MKL vs the average kernel SVM. Baseline mean AP 0.4434, mean AUC 0.8118.

mance differences. For ℓ^p-MKL vs the average kernel SVM, we have 10 statistically significant classes with 5 gains and 5 losses, while there are 12 statistically significant losses and only one gain in comparison between ℓ^1-MKL and the average kernel baseline.

Do sorted PHoCol features improve the BoW baseline? To answer this question we compared classifiers which takes both the BoW and PHoCol features with the baselines which rely only on the BoW representations. For each of the two cases and each class, we selected the best result in the AP score among various settings which will be explained later.

For combinations of the BoW and PHoCol features, we considered the six sets of base kernels in Table 1. For each set, the kernel weights are learned by ℓ^p-MKL with the tuning parameters $p \in \{1, 1.0625, 1.25, 1.5, 2\}$. The baselines only with the 6 BoW were also computed by taking the best result from ℓ^p-MKL and the average kernel SVM.

Table 1. The sets of base kernels tested.

set no.	BoWs	sorted PHoCols		
		color	sort costs	spatial tiling
1	all 6	opponent color 1 & 2	no	both 4×4 & 8×8
2	all 6	opponent color 1 & 2	yes	both 4×4 & 8×8
3	all 6	grey	no	both 4×4 & 8×8
4	all 6	grey	yes	both 4×4 & 8×8
5	all 6	normalized red, green, blue	no	both 4×4 & 8×8
6	all 6	normalized red, green, blue	yes	both 4×4 & 8×8

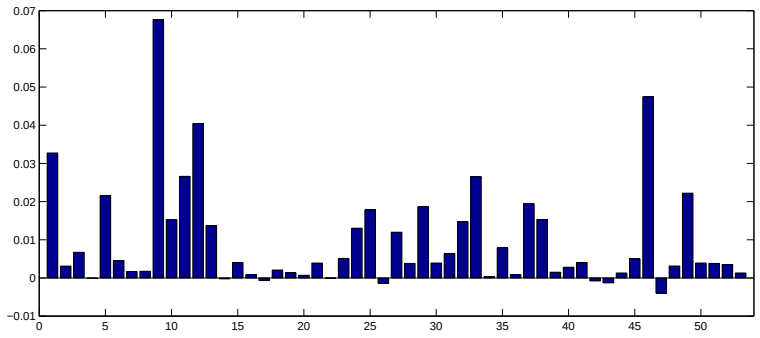

Fig. 3. Class-wise performance gains by the combination of the PHoCol and BoW over the standard BoW only. The baseline has mean AP 0.4434 and mean AUC 0.8118.

In Fig. 3, we can see several classes with larger improvements over the BoW baseline by employing the full setup including PHoCol features with the sort modification and the optimal kernel weights learned by ℓ^p-MKL. We also see slight decreases of the AP score on 6 classes out of all 53, where the worst setback is just of the size 0.004. In fact, they are rather minor compared to the large gains on their complement. Note that the combinations of PHoCol did not include the average kernel SVM as an option, while the best performances with the BoW only could be achieved by the average kernel SVM. Thanks to flexibility of ℓ^p-MKL, classification performances by the larger combination (PHoCol+BoW) were never much worse than the standard BoW classifiers, even PHoCols are much less informative.

The gains were statistically significant according to Wilcoxon signed rank test with the level $\alpha = 0.05$ on the 9 classes: Winter (13), Sky (21), Day (28), Sunset_Sunrise (32), Underexposed (38), Neutral_Illumination (39), Big_Group (46), No_Persons(47) and Aesthetic_Impression (51) in Fig. 3. This is not surprising, as we would expect for these outdoor classes to have a certain color

profile, while the two 'No' and 'Neutral' classes have a large number of samples for generalization via the learning algorithm.

We remark that the sorted PHoCol features are very fast to compute and that the MKL training times are negligible compared to those necessary for computing SIFT descriptors, clustering and assigning visual words. Actually we could compute the PHoCol kernels on the fly.

In summary, the result of this experiment shows that combining additional features with lower standalone performance can further improve recognition performances. In the next experiment we show that the non-sparse MKL is the key to the gain brought by the sorted PHoCol features.

Does averaging suffice for combining extra PHoCol features? We consider again the same two cases (i.e. PHoCol+BoW vs Bow only) as in the previous experiments. In the first case, the average kernels are always used as the combination of the base kernels in each set instead of ℓ^p-MKL and the best AP score was obtained for each class and each case. The performances of the second case were calculated in the same way as the last experiment.

From the upper panel of Fig. 4, we see a mixed result with more losses than gains. That is, the average kernels of PHoCol and BoW rather degrade the performance compared to the baselines with BoW only. Additionally, for the combination of PHoCol and BoW, we compared ℓ^p-MKL with the average kernel SVMs in the lower panel of Fig. 4. This result shows clearly that the average kernel fails in the combination of highly and poorly expressive features throughout most classes. We conclude that the non-sparse MKL techniques are essential to achieve further gains by combining extra sorted PHoCol features with the BoW representations.

Does the sort modification improve the PHoCol features? The default PHoCol features gave substantially better performances for the classes snow and desert on which the sorted ones do improve only somewhat compared to BoW models. We assume that the higher importance of color together with low spatial variability of color distributions in these concepts explains the gap. The default PHoCols without sorting degraded performances strongly in three other classes, where the sorted version does not lead to losses. In this sense, the sorting modification seems to make classifiers more stable on average over all classes.

4 Conclusions

We have shown that primitive color histograms can further enhance recognition performance over the standard procedure using BoW representations in most visual concepts of the ImageCLEF2009 photo annotation task, if they are combined optimally by the recently developed non-sparse MKL techniques. This fact was not known before and nobody has pursued this direction, because the average kernels constructed from such heterogenous features degrade classification

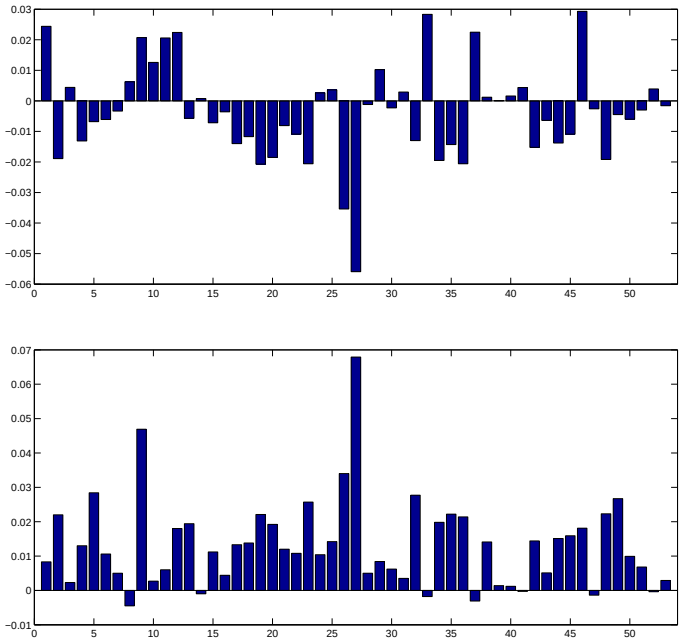

Fig. 4. Class-wise performance differences. Upper: combined Phocol and BoW by average kernel vs baseline with BoW only. Lower: combined Phocol and BoW by ℓ^p-MKL vs the same features with average kernel.

performance substantially due to high noise in the least informative kernels. Furthermore, we gave insights and evidences when ℓ^p-MKL is particularly useful: it can achieve better performance when combining informative and noisy features, even if the average kernel SVMs and the sparse ℓ^1-MKL fail.

Acknowledgements We like to thank Shinichi Nakajima, Marius Kloft, Ulf Brefeld and Klaus-Robert Müller for fruitful discussions. This work was supported in part by the Federal Ministry of Economics and Technology of Germany (BMWi) under the project THESEUS (01MQ07018).

References

1. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (1995)
2. Müller, K.R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B.: An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. **12**(2) (2001) 181–201
3. Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.R., Zien, A.: Efficient and accurate Lp-norm multiple kernel learning. In: Adv. in Neur. Inf. Proc. Sys. (NIPS). (2009)

4. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV '04, Prague, Czech Republic (May 2004) 1–22
5. Nowak, S., Dunker, P.: Overview of the CLEF 2009 large-scale visual concept detection and annotation task. In: Working Notes of CLEF 2009 Workshop. (2009)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comp. Vis. **60**(2) (2004) 91–110
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR '06, New York, USA (2006) 2169–2178
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pat. Anal. & Mach. Intel. **27**(10) (2005) 1615–1630
9. 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)