

# On the Benefits and the Limits of $\ell_p$ -norm Multiple Kernel Learning In Image Classification

Alexander Binder  
Technical University of Berlin  
Franklinstr. 28/29, 10587 Berlin, Germany  
alexander.binder@tu-berlin.de

Marius Kloft  
Technical University of Berlin  
mkloft@mail.tu-berlin.de

Wojciech Samek  
Technical University of Berlin  
wojwoj@mail.tu-berlin.de

Klaus-Robert Müller  
Technical University of Berlin  
klaus-robot.mueller@tu-berlin.de

Shinichi Nakajima  
NIKON Corporation  
Optical Research Laboratory, Tokyo, Japan  
nakajima.s@nikon.co.jp

Christina Müller  
Technical University of Berlin  
mueller@tu-berlin.de

Ulf Brefeld  
Yahoo! Research  
Barcelona, Spain  
brefeld@yahoo-inc.com

Motoaki Kawanabe  
Fraunhofer Institute FIRST  
Berlin, Germany  
motoaki.kawanabe@first.fraunhofer.de

## Abstract

*Many modern applications from the domain of image classification, such as natural photo categorization, come with highly variable concepts; to this end, state-of-the-art solutions employ a large number of heterogeneous image features, leaving a demand for combining information across many descriptors. In the paradigm of kernel-based learning, the multiple kernel learning (MKL) framework offers a principled way for learning linear combinations of feature groups / kernels — but the classical formulation is known to have limits in practice. We compare regular MKL with the recent  $\ell_p$ -norm multiple kernel learning methodology and the SVM using uniform kernel combination on VOC2009 and ImageCLEF2010 datasets using Bag-of-Words and simpler features. In our experiments we find advantages and shortcomings.*

## 1. Introduction

Combining different image representations to capture relevant traits of an image represents the state of the art in image classification. To this end, practitioners often resort to a uniform combination of kernels, which has been proven

to work well [4]. An alternative approach is multiple kernel learning (MKL) [6] that has been applied to image classification tasks using various image descriptors [3].

In this contribution, we study the limits and benefits of classical, sparse MKL and the recent  $\ell_p$  MKL [5], which outputs non-sparse kernel combinations, in object recognition tasks; we also compare our results to a simple SVM baseline using a uniform combination of kernels. We report on empirical results on image data sets from the PASCAL visual object classes (VOC) 2009 [2] and ImageCLEF2010 PhotoAnnotation [8] challenges. We remark that our focus is on a relative comparison of the methods so that we expect our analysis yielding similar conclusions when applied to alternative MKL methods such as MKL-KDA [10] or formulations using more optimization variables such as [1]. An extended version has been submitted to a journal.

## 2. Experiments

In our first computer vision experiments, we computed 32 kernels, all of them over varying color channels and spatial tilings in the spirit of [9], namely 15 Bag of Words (BoW) kernels over SIFT features, 8 BoW kernels over global color histograms, 4 kernels over color histograms, 5 kernels over global histograms of gradient orientations.

Table 3. Average AP scores on the VOC2009 test data set with class-wise selected  $\ell_p$ -norm by AP scores on the training set.

$\infty$	$\{1, \infty\}$	$\{1.125, 1.333, 2\}$	$\{1.125, 1.333, 2, \infty\}$	$\{1, 1.125, 1.333, 2\}$	all norms from the left
55.85	55.94	56.75	56.76	56.75	56.76

Table 4. Average AP scores on the ImageCLEF2010 training data set obtained by cross-validation with class-wise selected  $\ell_p$ -norm.

$\infty$	$\{1, \infty\}$	$\{1.125, 1.333, 2\}$	$\{1.125, 1.333, 2, \infty\}$	$\{1, 1.125, 1.333, 2\}$	all norms from the left
$39.11 \pm 6.68$	$39.33 \pm 6.71$	$39.70 \pm 6.80$	$39.74 \pm 6.85$	$39.82 \pm 6.82$	$39.85 \pm 6.88$

Table 1. Average AP scores attained on the VOC2009 test data. The average kernel is denoted as  $\ell_\infty$ .

norm	$\ell_1$	$\ell_{1.125}$	$\ell_{1.333}$	$\ell_2$	$\ell_\infty$ (SVM)
AP	54.58	56.43	<b>56.70</b>	56.34	55.85

Table 2. Average APs for ImageCLEF2010 data obtained by cross-validation. Standard deviations lie in 5.87 for  $\ell_1$  to 6.68 for  $\ell_\infty$

$\ell_1$	$\ell_{1.125}$	$\ell_{1.333}$	$\ell_2$	$\ell_\infty$ (SVM)
37.32	<b>39.51</b> $\pm 6.67$	39.48	39.13	$39.11 \pm 6.68$

From Tables 1 and 2, we conclude that non-sparse MKL outperforms sparse MKL as well as the average kernel when one has to select one method for all classes. Tables 3 and 4 are based on selecting the class-wise best classifier by cross-validation. Comparing the two rightmost entries in them shows that MKL methods can achieve good performance without relying on average kernel SVMs, however gains are limited in terms of AP score. The difference between ImageCLEF and VOC data is that sparse  $\ell_1$ -MKL can achieve gains in the former. We thus on one hand conclude from this experiment that by the use of a non-sparse  $\ell_{p>1}$  regularizer MKL can moderately help performance in image classification.

### 3. Discussion

As an interpretation of our results, we identify two arguments which favor average kernels and also one argument for learning kernel combinations.

#### 3.1. Randomness in BoW features

The first argument is the inherent randomness in the BoW kernel. To demonstrate this we recomputed the same BoW kernel ten times using k-means using initializations and the inherently randomized ERCF [7]. We can observe from Table 5 that averaging these kernels improve performance over the best single kernel despite using always the same SIFT features as inputs. Strikingly, ERCF which shows a higher amount of randomness also yields a higher gains from averaging. This indicates one reason why sparse methods may fail.

#### 3.2. Learning Kernels versus Prior Knowledge

The second argument is the trade-off between using prior knowledge and learning kernel combinations. A higher

Method	Best Single Kernel	Sum Kernel
VOC2009, k-Means	AP: $44.42 \pm 12.82$	<b>45.84</b> $\pm 12.94$
VOC2009, k-Means	Std: <b>30.81</b>	30.74
VOC2009, ERCF	AP: $42.60 \pm 12.50$	<b>47.49</b> $\pm 12.89$
VOC2009, ERCF	Std: <b>38.12</b>	37.89
ImageCLEF2010, k-Means	AP: $31.09 \pm 5.56$	<b>31.73</b> $\pm 5.57$
ImageCLEF2010, k-Means	Std: <b>30.51</b>	30.50
ImageCLEF2010, ERCF	AP: $29.91 \pm 5.39$	<b>32.77</b> $\pm 5.93$
ImageCLEF2010, ERCF	Std: <b>38.58</b>	38.10

Table 5. AP Scores and standard deviations showing amount of randomness in feature extraction: results from repeated computations of BoW Kernels with randomly initialized codebooks.

amount of prior knowledge reduces the potential of gains from learning methods. This can be shown by using a different kernel mixture with less BoW kernels which are known to be strong for VOC datasets. We observed for the latter a larger gap between  $\ell_{4/3}$ -MKL (AP: 52.33) and the average kernel SVM (AP: 50.33) but a lower best AP score in total compared to Table 1.

#### 3.3. Varying Informative Subsets Across Kernels

One argument for learning kernels is the hypothesis that typical computer vision kernels have subsets of data of varying size for which the kernel contains information. Using a method with a global kernel weight requires to weight the kernels accordingly. To analyze this we created synthetic data in which each kernel has a mutually disjoint subset of data  $n_k$  for which it is informative. A second experiment has been designed in a similar manner. Table 6 shows the results.

**Experimental Settings for Experiment 1 (3 kernels):**  
 $n_{k=1,2,3} = (300, 300, 500)$ ,  $p_+ := P(y = +1) = 0.25$ .  
 The features for the informative subset are drawn according to  
 $f_i^{(k)} \sim \begin{cases} N(0.0, \sigma_k) & \text{if } y_i = -1 \\ N(0.4, \sigma_k) & \text{if } y_i = +1 \end{cases}$ ,  $\sigma_k = \begin{cases} 0.3 & \text{if } k = 1, 2 \\ 0.4 & \text{if } k = 3 \end{cases}$ .  
 The features for the uninformative subset are drawn according to  
 $f^{(k)} \sim (1 - p_+)N(0.0, 0.5) + p_+N(0.4, 0.5)$ .

### References

[1] L. Cao, J. Luo, F. Liang, and T. S. Huang. Heterogeneous feature machines for visual recognition. In *JCCV*, pages 1095–1102, 2009.

Experiment	$\ell_\infty$ -MKL (SVM)	$\ell_{1.0625}$ -MKL	t-test p-value
1	68.72 $\pm$ 3.27	69.49 $\pm$ 3.17	0.000266
2	55.07 $\pm$ 2.86	56.39 $\pm$ 2.84	$4.7 \cdot 10^{-6}$

Table 6. Varying Informative Subsets of Data: AP Scores in Toy experiment using Kernels with disjoint informative subsets of data.

- [2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009), 2009. [1](#)
- [3] P. V. Gehler and S. Nowozin. Let the kernel figure it out; principled learning of pre-processing for kernel classifiers. In *CVPR*, pages 2836–2843, 2009. [1](#)
- [4] P. V. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, pages 221–228. IEEE, 2009. [1](#)
- [5] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011. [1](#)
- [6] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, pages 27–72, 2004. [1](#)
- [7] F. Moosmann, E. Nowak, and F. Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 30(9):1632–1646, sep 2008. [2](#)
- [8] S. Nowak and M. J. Huiskes. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010. [1](#)
- [9] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pat. Anal. & Mach. Intel.*, 2010. [1](#)
- [10] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler. Lp norm multiple kernel fisher discriminant analysis for object and image categorisation. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:3626–3632, 2010. [1](#)