

Augmented Kernel Matrix (AKM) vs Classifier Fusion (CF) for Object Recognition

Muhammad Awais, Fei Yan, Krystian Mikolajczyk, Josef Kittler
(m.rana,f.yan,k.mikolajczyk,j.kittler@surrey.ac.uk)

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK

Due to the importance of complementary information in feature combination [2, 3, 7], much research has been done in the field of feature design [4, 8] to diversify kernels. Kernels are often computed independently of each others thus can be highly informative, noisy or redundant. Therefore, Proper selection and fusion of kernels is crucial for optimizing the performance. The key idea of MKL, is to learn a linear combination of kernels by maximizing soft margin between classes [3]. Alternatively, AKM [10] is proposed arguing that in MKL a single kernel corresponding to a particular feature space is attributed a single weight. Therefore, MKL does not exploit information from individual samples in different feature spaces, e.g., in the context of object recognition, some samples can carry more shape information while others may carry more texture information for the same object category. In contrast to MKL, the main idea of CF is to construct a set of base classifiers and then classify a new sample by a weighted combination of their predictors. The success of CF especially AdaBoost [1] led to linear programming (LP) formulations of AdaBoost [6]. The fundamental problem with AKM is its large augmented matrix which makes it inapplicable to large datasets. We derive primal and dual of AKM and draw a comparison between AKM and CF, by carefully analysing the dual and the feature space of AKM.

We present a novel multiclass CF scheme (NLP-vMC) based on binary $v-LPBoost$ [6], which incorporates arbitrary norms $\{\ell_p, p \geq 1\}$ and optimizes the contribution from each class in each feature channel. We also incorporate nonlinear constraints in previously proposed binary $v-LPBoost$ and multiclass LPBoost [2] and show empirically that the nonlinear variants perform consistently better than their sparse counterparts, and baseline methods. We consider our extensive evaluation and comparison to the state-of-the-art fusion approaches as another important contribution of the paper. We perform experiments on multilabel and multiclass problems using standard benchmarks including Pascal VOC 2007, Flower 17, Flower 102 and Caltech101. Our multiclass formulation and nonlinear extensions of CF consistently outperforms the state-of-the-art MKL and sparse CF schemes. Note that we use object recognition datasets for evaluation, however, the proposed fusion schemes can be applied to any underlying pattern recognition problems provided that we have multiple feature channels.

Nonlinear Programming v-Multiclass (NLP-vMC): We consider one-vs-all formulation for multiclass case with N_C classes, i.e., response of each base classifier $g_r(x)$ now maps into an N_C dimensional space, $g_r(x) \mapsto \mathbb{R}^{N_C}$, and the output corresponding to c 'th class is denoted by $g_{r,c}(x)$. We extend the definition of margin for binary CF to multiclass CF as follows:

$$\rho(x_i, \beta) := \sum_{r=1}^n \beta_{(N_C(r-1)+y_i)} g_{r,y_i}(x_i) - \sum_{r=1}^n \sum_{j=1, j \neq i}^{N_C} \beta_{(N_C(r-1)+y_j)} g_{r,y_j}(x_i)$$

The classification confidence for examples x_i depends upon weight vector β and scores from base classifiers. The main difference between the two margins is that here, we are taking the class confidence of true target class and subtracts the combined effect of all the non-target classes from it, this difference is then summed over all feature channels. This is done for all n feature channels. The normalized (smallest) margin can then be defined as $\rho := \min_{1 \leq i \leq m} \rho(x_i, \beta)$. Inspired by the soft margin LP formulations of AdaBoost we propose to maximize the normalized margin ρ to learn linear combination of base classifiers. To avoid penalization of informative channels and to gain robustness against noisy feature channels, we change the regularization norm to handle any arbitrary norm $\ell_p, \forall p \geq 1$. The optimization problem is given by:

$$\max_{\beta, \xi, \rho} \rho - \frac{1}{vm} \sum_{i=1}^m \xi_i \quad (1)$$

$$\begin{aligned} s.t. \quad & \sum_{r=1}^n \beta_{(N_C(r-1)+y_i)} g_{r,y_i}(x_i) - \sum_{r=1}^n \sum_{j=1, j \neq i}^{N_C} \beta_{(N_C(r-1)+y_j)} g_{r,y_j}(x_i) \\ & \geq \rho - \xi_i \quad i = 1, \dots, m, \quad (2) \\ & \|\beta\|_p \leq 1, \rho \geq 0, \beta \geq 0 \quad \xi \geq 0 \quad \forall i = 1, \dots, m \end{aligned}$$

ML-Methods	1	$1+2^{-3}$	2	3	4	8
MKL	87.2±2.7	74.9±1.7	71.2±2.7	70.6±3.8	73.1±3.9	81.0±4.0
NLP- β	86.5±3.3	86.6±3.4	86.7±1.2	87.4±1.5	87.9±1.8	87.8±2.1
NLP-vMC	85.5±1.3	86.6±2.0	87.7±2.6	87.8±2.1	87.7±2.0	87.8±1.9
NLP-B	84.6±2.5	84.6±2.4	84.8±2.5	85.5±3.7	86.9±2.7	87.3±2.7

Comparison with State-of-the-Art	
MKL-prod [2]	85.5 ± 1.2
MKL-avg (ℓ_∞) [2]	84.9 ± 1.9
CF (ℓ_∞) / AKM	86.7 ± 2.7
CG-Boost [2]	84.8 ± 2.2
MKL (SILP or Simple) [2]	85.2 ± 1.5
LP- β [2]	85.5 ± 3.0
LP-B [2]	85.4 ± 2.4
MKL-FDA (ℓ_p)	86.7 ± 1.2

Table 1: Classification Rate on Flower17.

ML-Methods	1	$1+2^{-3}$	2	3	4	8	ℓ_∞
MKL	69.9	64.7	65.9	65.7	-	-	73.4
NLP- β	61.2	75.7	74.7	73.0	73.9	74.6	73.0
NLP-vMC	72.6	73.1	73.3	73.4	73.4	73.4	73.0
NLP-B	73.6	-	-	-	-	-	73.0

Comparison with State-of-the-Art	
MKL-prod	73.8
MKL-avg	73.4
MKL [5]	72.8

Table 2: Mean accuracy on Oxford Flower 102 dataset.

where $\frac{1}{vm}$ is the regularization constant and gives a trade-off between minimum classification confidence ρ and the margin errors. The main difference between this formulation and the binary CF formulation is the definition of margin used in the constraints in Eq. (2), in which the difference between the classification confidence of the true class and the joint confidence of all other classes is lower bounded. Note that the total number of constraints is equivalent to the number of training examples m plus one regularization constraint for l_p -norm (ignoring variables positivity constraints). Therefore, the difference in complexity, compared to the binary classifier fusion, is the increased number of variables in weight vector β , while having the same number of constraints. We also extend two multiclass CF schemes: LP- β and LP-B, proposed in [2] by introducing arbitrary regularization norms $\ell_p, \forall p \geq 1$, which avoids rejection of informative feature channels while being robust against noisy features channels. The optimization problems of NLP-vMC, NLP- β and NLP-B, are nonlinear separable convex and can be solved using MOSEK.

We present results on a multi-label dataset, namely, Pascal VOC 2007 and three multiclass datasets, namely, Flower17, Flower102 and Caltech101. We show the results for only two datasets due to lack of space other datasets have similar results. The proposed methods perform better than the state-of-the-art MKL methods. In addition to this, the non-sparse version of CF is performing better than sparse selection of feature channels. **Acknowledgements.** This research was supported by UK EPSRC EP/F003420/1, EP/F069421/1 and the BBC R&D grants.

- [1] Y. Freund and R. Schapire. A Desicion-Theoretic Generalization of On-Line Learning and an Application to Boosting. In *CLT*, 1995.
- [2] P. Gehler and S. Nowozin. On Feature Combination for Multiclass Object Classification. In *ICCV*, 2009.
- [3] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *JMLR*, 5:27–72, 2004.
- [4] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.
- [5] M-E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *ICCVGIP*, 2008.
- [6] G. Rätsch, B. Schölkopf, A. Smola, S. Mika, K.R. Müller, and T. Onoda. Robust Ensemble Learning for Data Analysis. In *PACKDDM*, 2000.
- [7] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [8] K. van de Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *CVPR*, 2008.
- [10] F. Yan, K. Mikolajczyk, J. Kittler, and A. Tahir. Combining Multiple Kernels by Augmenting the Kernel Matrix. In *International Workshop on Multiple Classifier Systems*, 2010.