AN EFFICIENT SYSTEM FOR COMBINING COMPLEMENTARY KERNELS IN COMPLEX VISUAL CATEGORIZATION TASKS

David Picard, Nicolas Thome and Matthieu Cord

LIP6 UPMC Paris 6
4 place Jussieu
75005 Paris France

ABSTRACT

Recently, increasing interest has been brought to improve image categorization performances by combining multiple descriptors. However, very few approaches have been proposed for combining features based on complementary aspects, and evaluating the performances in realistic databases. In this paper, we tackle the problem of combining different feature types (edge and color), and evaluate the performance gain in the very challenging VOC 2009 benchmark. Our contribution is three-fold. First, we propose new local color descriptors, unifying edge and color feature extraction into the *"Bag Of Word"* model. Second, we improve the Spatial Pyramid Matching (SPM) scheme for better incorporating spatial information into the similarity measurement. Last but not least, we propose a new combination strategy based on ℓ_1 Multiple Kernel Learning (MKL) that simultaneously learns individual kernel parameters and the kernel combination. Experiments prove the relevance of the proposed approach, which outperforms baseline combination methods while being computationally effective.

1. INTRODUCTION

Image categorization consists in predicting, in a given image, the presence/absence of an example of a pre-defined class. Combining descriptors for improving categorization performances has extensively been studied in the last decade. Multiple Kernel Learning (MKL) is appealing for that purpose, since it offers the possibility to jointly learn the weighting of the different channels and the classification function. Some recent MKL studies [1, 2] provide evaluations in complex datasets such as VOC challenge [3]. However, they use redundant descriptors (various SIFT or HoG variants), that only capture a single image modality (edge). Other approaches [4] use complementary features, but do not provide evaluations on such challenging databases.

In this paper, we consider the problem of learning combinations of complementary descriptors to make the categorization task more efficient. The proposed approach is schemed in figure 1. There are three main areas of novelty: extracting complementary informative descriptors (section 2.1), building an image representation from each feature (section 2.2), and learning a category-specific combination between them so that improving categorization performances (section 2.3).

2. PROPOSED METHOD

2.1. Unified extraction of local descriptors

In this paper, we propose to represent each image modality by a *"Bag Of Word"* model [5]. Regarding image categorization, this representation proves to reach state of the art performances when dealing with complex images , and significantly outperforms global image signatures. In this paper, we propose to apply the *"Bag Of Word"* model for all extracted descriptors. We focus here on edge and color features, but the method naturally extends to other descriptors (*e.g.* texture).

We choose to extract the different descriptors on a regular grid over the image (dense sampling point strategy). Around each patch with its associated scale, a set of descriptors is computed over a local neighborhood. An offline clustering strategy (*k-means*) is performed to compute a visual codebook. In this paper, we use $K = 4000$ visual words for all descriptors. Each image modality is thus represented by a histogram of visual words, with incorporated spatial information (section 2.2). The most successful strategy for computing the histograms is a "semi-soft assignment" that assigns for each descriptor a vote of 1 to its $N = 10$ closest visual words [6].

2.1.1. Edge descriptors

We used opponentcolorSIFTs (oc-SIFTs) [7] as local edge descriptors, a variant the well known SIFT descriptors [8]. oc-SIFT histogram is a concatenation of three 1-D SIFT histograms based on the channels of the opponent color space O_1, O_2, O_3:

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{3} \end{pmatrix} \quad (1)$$

where R, G, B refers to the standard color space. oc-SIFT leads thus to a descriptor of dimension $128 \times 3 = 384$. Due to the SIFT normalization, oc-SIFTs are invariant to changes in light intensity (see [7]).

2.1.2. Color descriptors

We choose the Hue Saturation Value (HSV) space for extracting a local color descriptor, since HSV is known to be perceptually appropriate. Thus, at each grid position with its associated scale, we compute a color histogram over the region using a quantization of the HSV space. We quantify H with 8 bins, S with 6 bins, and V with 3 bins. This leads to a descriptor of dimension $8 \times 6 \times 3 = 144$. Note that some recent works [7] propose to extract local color descriptors and to build visual dictionaries from them. However, all proposed color descriptors (*e.g.* RGB histograms) consider the different color axes independently by concatenating 1-d histograms. We claim that our representation, a 3d histogram, is a more powerful description of each patch by its capacity to encode correlations between axes.

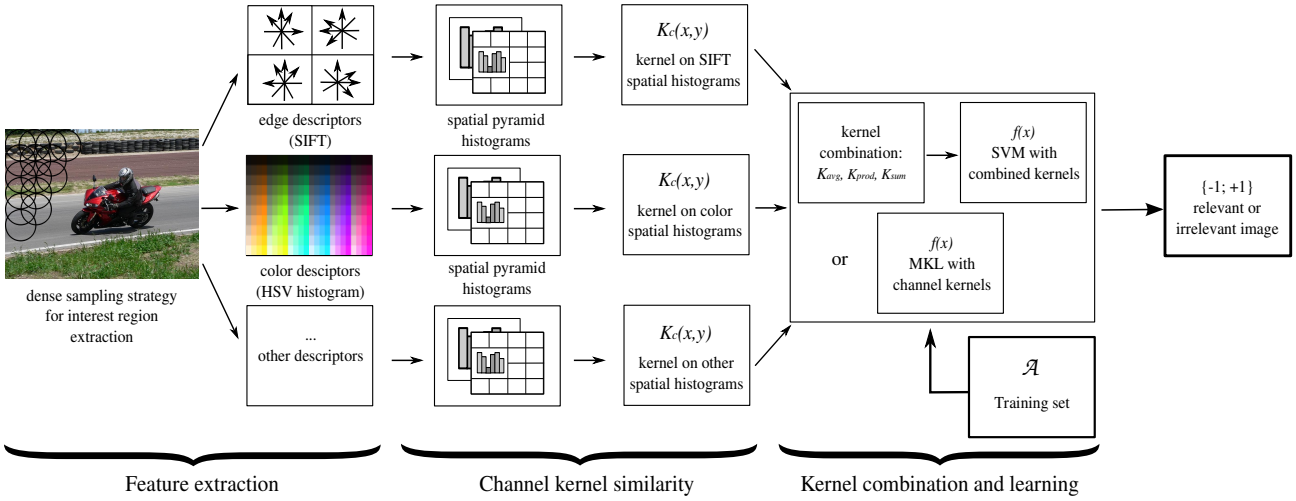


Fig. 1. Approach Overview.

2.2. Channel kernel similarity: adapted SPM scheme

In order to take into account the spatial information for each extracted image descriptor, we propose to adapt the *"Spatial Pyramidal Matching"* (*SPM*) introduced in [9]. The *SPM* works as follows: at a specific scale l, the image is divided following a regular grid, and a histogram of visual words is computed on each resulting region. In the last PASCAL VOC Challenge [3], the most common scheme of partitioning the image used three decomposition scales: a histogram over the entire image (scale 0), a 2×2 grid resulting in 4 histograms (scale 1), and a 3×1 grid resulting in 3 horizontal band histograms (scale 2). We propose here a decomposition strategy with two scales, keeping the global histogram at scale 0. At scale 1, however, we use a decomposition of 3×3 overlapping windows. Each window corresponds to a quarter of the image, with an overlap of half the window size. For each of the 10 regions r previously defined and for each descriptor channel c, we defined the similarity $k_{c,r}(\mathbf{x}_i, \mathbf{x}_j)$ between two images $\mathbf{x}_i$ and $\mathbf{x}_j$ as the following Gaussian kernel:

$$k_{c,r}(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma_c \, d_{\chi^2}\left(\mathbf{x}_i^{(c,r)}; \, \mathbf{x}_j^{(c,r)}\right)} \tag{2}$$

Where $\mathbf{x}^{(c,r)}$ denotes the histogram of visual words for $\mathbf{x}$ associated with descriptor c and region r, and $d_{\chi^2}(\cdot, \cdot)$ is the χ^2 distance. The similarity $K_c(\mathbf{x}_i, \mathbf{x}_j)$, for channel c, between images $\mathbf{x}_i$ and $\mathbf{x}_j$ is thus defined as a weighted linear combination of local similarities $k_{c,r}(\mathbf{x}_i, \mathbf{x}_j)$, as in [9]:

$$K_c(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{10} w_r \; k_{c,r}(\mathbf{x}_i, \mathbf{x}_j) \tag{3}$$

where w_r weights are set proportional to the size of the considered regions.

2.3. Combining the kernels

Different strategies can be carried out to perform the combination of different features types.

2.3.1. Baseline early fusion - product kernel

The simplest way of combining N_c channels is to merge the corresponding signatures into a single vector by concatenating the associated feature spaces. This approach is called *early fusion*. In our SPM context, the early fusion is to be done for each histogram of the pyramid. The resulting major kernel is the following:

$$K_{prod}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{r=1}^{10} w_r e^{-\sum_{c=1}^{N_c} \gamma_c \, d_{\chi^2}\left(\mathbf{x}_i^{(c,r)}; \, \mathbf{x}_j^{(c,r)}\right)} \tag{4}$$

$K_{prod}(\mathbf{x}_i, \mathbf{x}_j)$ can be further decomposed as follows: $K_{prod}(\mathbf{x}_i, \mathbf{x}_j) = \sum_r w_r \prod_c e^{-\gamma_c \, d_{\chi^2}\left(\mathbf{x}_i^{(c,r)}; \, \mathbf{x}_j^{(c,r)}\right)} = \sum_r w_r \prod_c k_{c,r}(\mathbf{x}_i, \mathbf{x}_j)$.

Hence, an early fusion scheme with Gaussian kernels is equivalent to a product of kernels in each pyramid region.

2.3.2. Baseline intermediate fusion - weighted sum kernel

An other way of combining the N_c different channel specific kernels is to compute a weighted sum over them:

$$K_{sum}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{c=1}^{N_c} \beta_c K_c(\mathbf{x}_i, \mathbf{x}_j) = \sum_{c=1}^{N_c} \beta_c \sum_{r=1}^{10} w_r k_{c,r}(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

If $\beta_c = \frac{1}{N_c} \; \forall c$, K_{sum} is a simple averaging kernel K_{avg}. Otherwise, β_c can be further optimized using a cross-validation procedure. However, this brute-force strategy rapidly becomes intractable when N_c increases.

Using the explicit inner product of induced space, we have $k_{c,r}(\mathbf{x}_i, \mathbf{x}_j) = < \phi(\mathbf{x}_i^{(c,r)}), \phi_c(\mathbf{x}_j^{(c,r)}) >$. Therefore, $K_{sum}(\mathbf{x}_i, \mathbf{x}_j) = \sum_r w_r \sum_c \beta_c < \phi_c(\mathbf{x}_i^{(c,r)}), \phi_c(\mathbf{x}_j^{(c,r)}) >$. The weighted sum combination can thus be interpreted as an intermediate fusion of channels, by the concatenation of the signatures in the induced space.

2.3.3. Learning a non-sparse combination via ℓ_1 MKL

The problem of learning a linear combination of different kernels has recently been formalized as Multiple Kernel Learning [10]. In this case, the combination weights β and the support vector weights α are learned together in a joint optimization. The goal is to find the optimal classification function f defined as follows:

$$f(\mathbf{x}) = \sum_i \alpha_i y_i \sum_m \beta_m k_m(\mathbf{x}, \mathbf{x}_i) - b \tag{6}$$

Recent works attempting at using MKL on image dataset for combining different channels [1, 4] use MKL optimization algorithms based on ℓ_1 norm to regularize the kernel weights, like SimpleMKL [11]. Since this leads to sparse solutions, most studies report that MKL is often outperformed by simple baseline methods

(product or averaging) when dealing with complementary and informative kernels [1, 4]. In our case, we do not want to select either the SIFT or the color channel, but we aim at finding a proper weighting between them. Except [2], very few approaches use ℓ_2 MKL optimization schemes to find a non-sparse combination of complementary descriptors.

In this paper, we propose a hybrid strategy that does not lead to ignore informative image modalities during training. Importantly, our algorithm learns individual kernel parameters (γ in the Gaussian case) and the kernel combination coefficients (β_m) simultaneously. Other approaches like baseline methods (section 2.3.1) or ℓ_2 MKL use a two-step procedure: optimal γ is first determined by cross-validation, and combining the kernels is then performed for a fixed γ. This leads to a sub-optimal parameter estimation with respect to our global optimization scheme. As we verify experimentally (section 3), the proposed algorithm leads to an accurate parameter estimation while being computationally efficient. Thus, for each channel c, we form a set of M kernels $K_{c,\gamma}$, and use a ℓ_1 MKL strategy to select the relevant γ parameter. The sparse solution output by ℓ_1 MKL algorithms is therefore used as an option to cross-validation. (see [11]). Our adapted MKL problem formulation leads to find the optimal function of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{N_e} \alpha_i y_i \sum_{c=1}^{N_c} \sum_{\gamma=\gamma_1}^{\gamma_M} \beta_{c,\gamma} k_{c,\gamma}(\mathbf{x}, \mathbf{x}_i) - b \qquad (7)$$

where the joint optimization is performed on α_i (N_e parameters) and $\beta_{c,\gamma}$ ($N_c \times M$ parameters).

3. EXPERIMENTS AND RESULTS

We evaluate the proposed approach on the PASCAL VOC2009 challenge. This dataset is now accepted as being the most publicly available difficult benchmark for object image classification, due to multiple object viewpoints, large scale range, complex background, *etc*.

3.1. Setup

We used the *train* subset for training the classifier (about 3500 images) and the *val* subset for testing (about 3500 images). We evaluate the categorization performances for the individual edge and color descriptors, and different combination methods: baseline methods, (averaging & product), weighted sum kernel, and our adapted MKL approach. Not that except for MKL, the γ parameters for the individual kernels are optimized individually on the *val* subset. Table 1 presents the Average Precision (AP) for different combination strategies, and for the 15 VOC categories where adding color increases performances (for the 5 remaining categories, performances are similar to the use of the SIFT descriptor alone).

3.2. Evaluation of descriptors and baseline combinations

Regarding individual descriptor performances, the approach only using oc-SIFT is the run we submitted for the VOC2009 Challenge, ranked 6 among 20 engaged groups. Thus, oc-SIFT alone not surprisingly performs far better than the color descriptor (MAP of 47.8% *vs* 29.5%, column 1 and 2 of table 1). Indeed, in such object image database, edge is a much more powerful feature than color. However, we want to stress that a random classifier would achieve a MAP around 7%: our proposed color descriptor contains thus a significant amount of information for the categorization task.

Contrarily to recent works [1, 4], we notice that combining edge and color using baseline methods gives very disappointing results: MAP of 45.8% & 47.5% for product & averaging, respectively (column 3 and 4 of table 1). These combination strategies yield thus performances worse than using only oc-SIFT. In [1], redundant edge descriptors are used whereas complementary features are merged in [4]. However, in both cases, the different descriptors have comparable categorization abilities. Therefore, baseline methods seem inappropriate for combining complementary descriptors with significant variations regarding categorization performances.

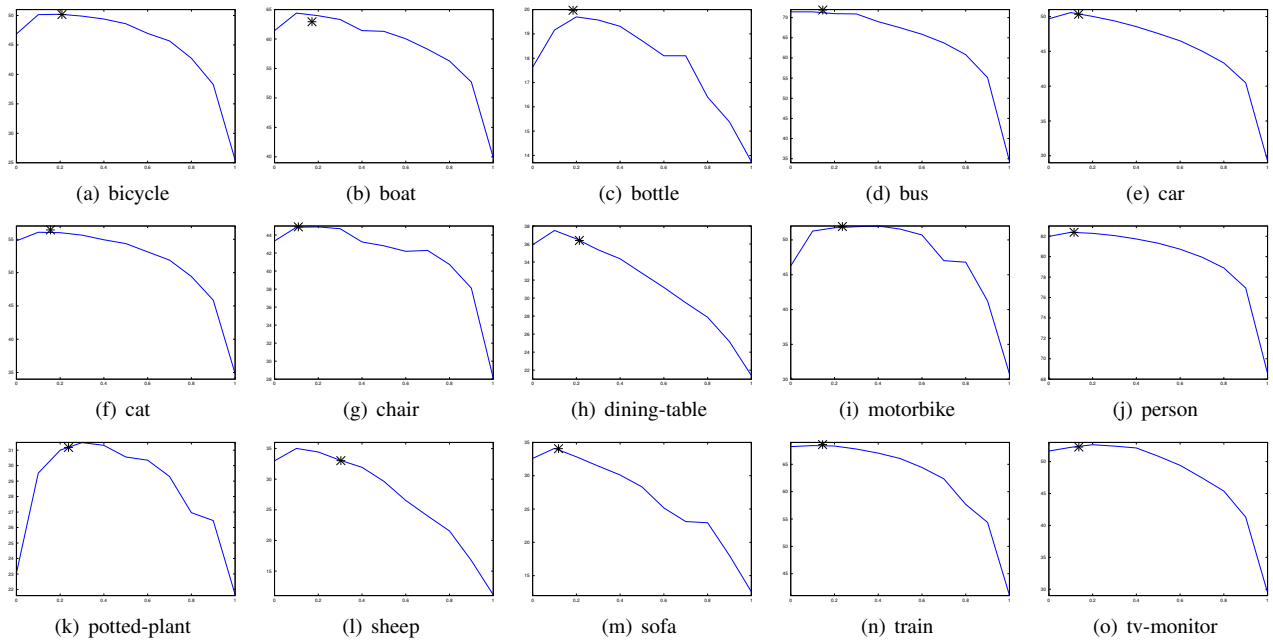
category	SIFT	Color	Prod	Avg	BS	MKL
bicycle	46.9	25.5	44.7	48.6	**50.2**	**50.2**
boat	61.4	39.9	59.7	61.3	**64.4**	62.9
bottle	17.6	13.7	19.7	18.7	19.7	**20.0**
bus	71.4	34.3	65.5	67.5	71.4	**71.9**
car	49.7	29.2	48.0	47.6	**50.6**	50.3
cat	54.8	34.8	50.9	54.4	56.1	**56.4**
chair	43.3	28.1	42.9	42.8	**44.9**	**44.9**
dining-table	35.9	21.4	33.6	32.8	**37.5**	36.4
motorbike	46.3	30.8	47.9	51.6	**52.0**	51.9
person	82.0	68.6	80.0	81.3	**82.4**	**82.4**
potted-plant	23.0	21.7	27.8	30.6	**31.5**	31.2
sheep	33.0	11.1	27.5	29.7	**35.0**	33.0
sofa	32.6	12.6	26.8	28.3	**34.1**	**34.1**
train	68.2	41.1	63.5	66.1	68.4	**68.6**
tv-monitor	51.6	29.5	48.1	50.8	**52.6**	52.3
MAP	47.8	29.5	45.8	47.5	**50.1**	49.8

Table 1. VOC 2009: (M)AP for different combination strategies.

3.3. Evaluation of learning-based combination strategies

The performances of the weighted sum kernel (equation 5) depend on the weights β_i between color and edge kernels. Column 5 of table 1 presents results of the best weighting, learned by cross-validation, that we denote BS ("Best Sum"). BS almost always performed better than the best performing kernel (oc-SIFT): the mean gain is about 2.3% (MAP from 47.8% to 50.1%). This confirms the soundness of combining edge and colors descriptors using a weighted sum kernel. Indeed, although color descriptor performance is limited, this proves that a category-specific weighting with the edge descriptor can significantly improve performances. On the category *potted-plant*, the gain reaches more than 8% of MAP, which corresponds to more than 30% improvement. Let alone for the categories *bicycle*, *boat*, *bottle*, *cat*, *dinning-table*, *motorbike*, *potted-plant* and *sofa*, the mean gain is about 4% of MAP.

The main drawback of the weighted sum kernel is its computational cost. Indeed, it requires to first determine γ for each channel, and then the weights β of the combination. This brute-force strategy is feasible for few channel kernels (here $N_c = 2$), but rapidly becomes intractable when N_c increases. Our adapted MKL approach (section 2.3.3) offers an elegant alternative to this problem by jointly learning γ and β. We use 5 kernels with variable γ for each channel, leading to a 10 kernel weighted sum. The learning is performed using the SimpleMKL algorithm [11], and is very fast: in our experiments, the gradient based optimization always converges very quickly, and learning for a category can be performed in a few minutes. Categorization performances using MKL are shown in the last column of table 1. As we can see, MKL gives results similar to BS kernel, leading to a MAP of 49.8%. Therefore, Contrarily to [1, 4], we notice that MKL significantly outperforms baseline methods (product and averaging). This is because the ℓ_1 algorithm



(a) bicycle (b) boat (c) bottle (d) bus (e) car

(f) cat (g) chair (h) dining-table (i) motorbike (j) person

(k) potted-plant (l) sheep (m) sofa (n) train (o) tv-monitor

Fig. 2. Analysis of the weighting regarding performances of the weighted sum kernel. Each curve represent Average Precision for each category, versus the weight β_{col} of the color kernel. $\beta_{col} = 0$ corresponds to using oc-SIFT only, whereas $\beta_{col} = 1$ corresponds to using color only. The mark denotes AP for the proposed MKL approach, with respect to the sum of weights corresponding to the color kernels.

directly used in [1, 4] with complementary kernels leads to a spare solution that discard informative channels from the combination. At the opposite, our approach keeps all informative image modalities for performing the combination.

Figure 2 shows the influence of the weights for the sum combination. Let us denote β_{col} the weight of the color kernel, and $(1-\beta_{col})$ the weight of the SIFT kernel. The MAP is shown against β_{col}, normalized to the value obtained for $\beta_{col} = 0$, for each category. As we can see, the optimal combination is dependent on the category, but always resides in low β_{col} values, except for *potted-plant*, for which the color descriptors were almost as good as the SIFT descriptors. We also draw a mark corresponding to the MAP of the MKL method against the weights obtained by the color kernels after the optimization. We can notice that MKL selects a β_{col} parameters close to the optimal value determined by BS. Note that regarding β learning, BS represents an "ideal" kernel since the parameters are determined by maximizing AP on the *val* subset. In that sense, the performances for BS (and identically for other individual descriptor kernels and baseline combination methods regarding γ) are over-estimated with respect to MKL. Therefore, the fact that MKL reaches performances close to BS illustrates the effectiveness of the conjoint learning of β and γ parameters, that is better-founded than the two-step optimization used with the other approaches. For four categories (see table 1), MKL outperforms BS, proving the efficiency of the joint learning.

4. CONCLUSION

In this paper, we propose an efficient strategy for combining complementary features using the kernel framework. The kernel combination is based on an adaptation of a ℓ_1 MKL algorithm that do not lead to ignore any informative image modality during learning. We evaluate the proposed learning-based approach on the challenging VOC'09 dataset, and show that it significantly outperforms baseline combination methods. Moreover, our algorithm is computationally effective, and provides a promising alternative to cross-validation by its capacity to simultaneously learn individual kernel parameters and the kernel combination.

5. REFERENCES

[1] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009.

[2] Fei Yan, Krystian Mikolajczyk, Josef Kittler, and Muhammad Tahir, "A comparison of l1 norm and l2 norm multiple kernel svms in image and video classification," *CBMI, International Workshop on*, vol. 0, pp. 7–12, 2009.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results," http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[4] Peter V. Gehler and Sebastian Nowozin, "On feature combination for multiclass object classification," in *IEEE ICCV*, 2009.

[5] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, Oct. 2003, vol. 2, pp. 1470–1477.

[6] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Trans. PAMI*, vol. in press, 2009.

[7] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on PAMI*, , no. in press, 2010.

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *IJCV*, 2003, vol. 20, pp. 91–110.

[9] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR '06*, Washington, DC, USA, 2006, pp. 2169–2178, IEEE Computer Society.

[10] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *ICML '04*, 2004, p. 6.

[11] Alain Rakotomamonjy, Francis Bach, Stephane Canu, and Yves Grandvalet, "SimpleMKL," *JMLR*, vol. 9, pp. 2491–2521, 2008.