A New Multiple Kernel Approach for Visual Concept Learning

Jingjing Yang[1,2,3], Yuanning Li[1,2,3], Yonghong Tian[3], Lingyu Duan[3], and Wen Gao[1,2,3]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
[2] Graduate School, Chinese Academy of Sciences, Beijing, 100039, China
[3] The Institute of Digital Media, School of EE & CS, Peking University, Beijing, 100871, China
{jjyang,ynli}@jdl.ac.cn, {yhtian,lingyu,wgao}@pku.edu.cn

Abstract. In this paper, we present a novel multiple kernel method to learn the optimal classification function for visual concept. Although many carefully designed kernels have been proposed in the literature to measure the visual similarity, few works have been done on how these kernels really affect the learning performance. We propose a Per-Sample Based Multiple Kernel Learning method (PS-MKL) to investigate the discriminative power of each training sample in different basic kernel spaces. The optimal, sample-specific kernel is learned as a linear combination of a set of basic kernels, which leads to a convex optimization problem with a unique global optimum. As illustrated in the experiments on the Caltech 101 and the Wikipedia MM dataset, the proposed PS-MKL outperforms the traditional Multiple Kernel Learning methods (MKL) and achieves comparable results with the state-of-the-art methods of learning visual concepts.

Keywords: Visual Concept Learning, Support Vector Machine, Multiple Kernel Learning.

1 Introduction

With the explosive growth of images, content-based image retrieval (CBIR), which searches images whose low-level visual features (e.g., color, texture, shape, etc.) are similar to those of the query image, has been an active research area in the last decade. However, retrieving images via low-level features has been proven to be unsatisfactory since low-level visual features cannot represent the high-level semantic content of images. To address this so-called semantic gap issue [1], a variety of machine learning techniques have been used to map the image features to semantic concepts, such as scenes (e.g. indoor/outdoor [2], and some specified natural scenes [3]) and object categories (e.g. airplane/motorbike/face) [4,5]. Most of these methods follow a supervised learning scheme, which learns visual concepts from a set of manually labeled images and classifies unseen images into one of learned concepts. However, these methods are application-specific and hard to be generalized from a relatively small data set to a much larger one.

B. Huet et al. (Eds.): MMM 2009, LNCS 5371, pp. 250–262, 2009.

Fig. 1. Samples of "airplane" in Caltech101

Fig. 2. Samples of "military aircraft" in Wikipedia

Learning visual concept from numerous images is difficult: concept within the images has visual ambiguities. On one hand, images of different concepts can exhibit somewhat similarity on different attributes: scale, shape, color, texture, etc. For objects such as cars and buses, shape is an important clue to discriminate them while color is usually not. On the other hand, the same object can yield different visual appearance due to the variety of view point, illuminance, or shelter. Different instances of the same concept can also have diversity in appearance (see Fig.1, 2). In short, instances of visual concepts are usually of variation and redundancy in multiply image feature spaces. Learning visual concept requires making a trade-off between the invariant and discriminative power.

In this paper, we present a per-sample based multiple kernel learning (PS-MKL) method for visual concept, which aims at decoding the discriminative ability of every training sample in multiple basic kernel spaces and forms a unique and general classification function for each visual concept. In addition to learning an important weight for each training sample, we also determine a sample specific linear combination of basic kernels. We show that the learning of the sample weights and the per-sample based kernel weights can be formed as a Max-Min problem and solved in a global optimal manner.

We apply our technique on two dataset, the Caltech101 and the Wikipedia MM dataset [6]. Caltech101 is introduced in 2004 [7] with a numerous object categories. However, there is little variation in pose or scale within the object class (see Fig.1). To this end, we go on experiments on the Wikipedia image collection which contains approximately 150,000 images that are of diverse topics and more similar to those encountered on the Web (see Fig.2). We highlight the main contributions of our work as follow:

1. We introduce PS-MKL for visual concept. This technique provides a tractable solution to the combinational sample weighting and the sample-level kernel weighting, which is thinner than concept level in previous works [8,9].
2. Multiple kernels are automatically selected for each weighted training sample. New kernels can be utilized easily and systematically assess their performances in PS-MKL.

The remainder of this paper is organized as follows. Sec. 2 reviews the related work. In Sec.3, we propose the PS-MKL framework as a classifier for learning visual concept. Details of the learning procedure are described in Sec. 4. We depict the application of object recognition and image retrieval and present the experimental results in Sec. 5. The conclusions and future work are discussed in Sec. 6.

2 Related Works

Learning visual concept within the image is currently one of most interesting and difficult problems in computer vision. Much progress has been made in the past

decades in investigating approaches that capture the visual and geometrical statistics of visual concepts. Three types of related approaches: generative approach, exemplar based approach, and kernel approach, will be surveyed in this section.

Generative Approach in Visual Concept Learning
In the past few years, generative approaches have become prevalent in visual concept learning. They learn concept from a share of low level features, and introduce a set of latent variables to fuse various cues. For example, part-based method (e.g. constellation mode [9,10]) and bag of words method [5,11] learn object categories via shape invariance. In recognition stage, a generative approach estimates the joint probability density function with Bayes' rule. Ng and Jordan [12] demonstrated both analytically and experimentally that in a 2-class setting the generative approach often has better performance for small numbers of training examples. However, recognition requires intermediate results based on the pre-computed and shared low level features. The performance declines sharply with the scale of the learning concepts.

Exemplar Based Method for Visual Concept Learning
Some researchers solve visual concepts learning problem in a manner of data association, which uses the image nearest neighbor to infer its own identity. Works in [13] learn separate distance function for each exemplar to improve recognition. Recently, several systems show that k-nearest-neighbor (KNN) search based on appearance archive surprising results [14].

Exemplars in these methods are assumed to be independent with each other and treated equally. Although exemplar based method are efficient in training stage, recognition is time consuming to match test image with every training exemplar. Another requirement is a very large data set to obtain all possible configurations of the visual concept within the image.

Kernel Based Method for Visual Concept Learning
While generative approach and exemplar based approach have archived some success, kernel based discriminative methods are known to efficiently find decision boundaries in kernel space and generalize well on unseen data [15].

(a). Single Kernel Method for Visual Concept Learning: So far, various kernels, carefully designed to capture different types of clues for visual concept learning, have been employed in Support Vector Machine (SVM) to find the optimal separating hyper-plane between the positive and negative classes. Kernels based on multi-resolution histogram, which compute image global histogram, are proposed in [16] to measure the image similarity at different granularity. Pyramid matching kernel which matches images with local spatial coordinates is proposed in [17] to enforce loose spatial information. Kernels for local feature distribution are presented in [18] to capture the image local context. However, these methods are designed to operate on one fixed input of feature vector, where each vector or vector entry corresponds to a particular aspect of image. These methods are not directly applicable to large scale concept learning since different attributes of the image or visual concepts are treated equally so that attributes can be easily overwhelmed by the major one. As mentioned

in Sec. 1, weighting of multiple aspects is needed to be conducted to improve the discriminative power for visual ambiguity.

(b). Multiple Kernel Method for Visual Concept Learning: Recently, much progress has been made in the field of multiple kernel learning [19,20]. These methods follow a same framework as a linear combination of basic kernels but differ in the cost function to optimize. The combination of basic kernels shows two benefits. First, there is no need to work on a combined high dimensional feature space. Second, different types of feature, such as shape, color, texture can be formulated effectively in a uniform formula to avoid the over-fitting problem. However, weights of basic kernels in these works are learned at concepts level, where personalities of different concept instances are ignored.

Considering the approaches and the corresponding problems mentioned above, we propose a new multiple kernels learning method PS-MKL for visual concept to achieve a balance between the invariant and discriminative power throughout the training samples. In PS-MKL, sample selection is combined with a kernel weighting at sample level which is a thinner granularity over previous methods. Both the sample weights and the kernel weights are optimized in a unified convex problem.

3 A New Multiple Kernel Learning Framework

Given a labeled image dataset as $D_l = \{x_i, y_i\}_{i=1}^{N}$, where x_i is the i_{th} labeled training examples and $y_i = \{\pm 1\}$ is the corresponding binary label for a given visual concept, our goal is to train a classifier $f(x)$ from the training samples, which can accurately classify the unlabeled image dataset D_u.

The flowcharts of identifying visual concept of an unseen image by three binary classifiers are shown in Fig.3. All of the three strategies adopt a similar kernel based framework but different kernel structures. These top-down flowcharts have three layers including input layer, middle layer and decision layer. In the input layer on the top, x represents a test sample which needs to be identified with the relevant visual concept. In the middle layer, similarities between test sample x and training samples $\{x_i\}_{i=1}^{N}$ are measured respectively via different kernel structures. In the last layer, the sign of decision function including linear combination of weighted kernels determines the result whether the test sample x contain the given concept.

The leftmost flowchart of Fig.3 depicts the first paradigm: standard SVM method, which employs a single kernel to measure the sample similarity. The center flowchart of Fig.3 shows the second paradigm: multiple kernel method, which compares the similarity of each basic kernel is related with the corresponding feature sub-space and the adoptive kernel function. The rightmost flowchart of Fig.3 displays the third paradigm: our proposed per-sample based multiple kernel method, which also utilizes multiple basic kernels to measure similarity. The difference is that combined modality for basic kernels is distinguishable among different training sample. The visual characteristics of each training sample make an impact on the weight of each basic

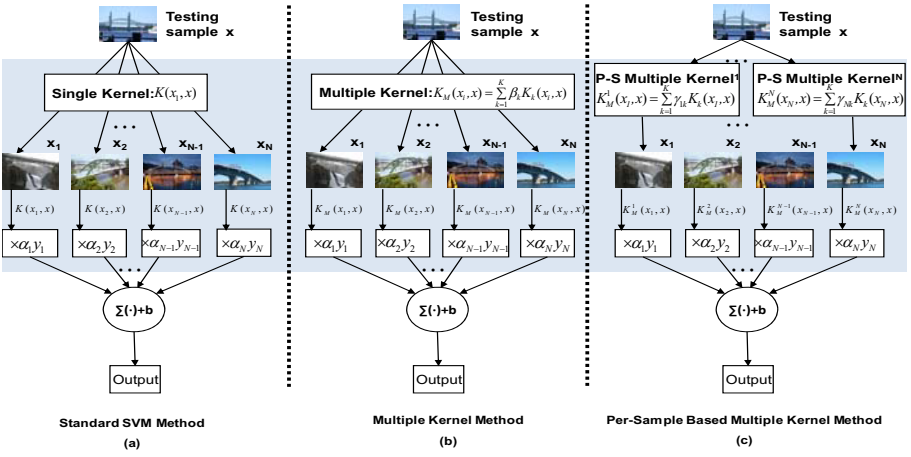


Fig. 3. Three paradigms of learning image concepts using (a) Standard SVM Method; (b) Multiple Kernel Method; (c) Per-Sample Based Multiple Kernel Method

kernel, which results in more learning parameters in such classifier compared with the former two strategies.

In this section, we first briefly review the standard SVMs and the traditional multiple kernel method in sec.3.1 and sec.3.2 respectively. Inspired by these methods, our proposed per-sample based multiple kernel method for visual concept learning is introduced in Sec.3.3.

3.1 Standard SVM

SVM uses a feature map ϕ to project the original data in input space to a higher dimensional feature space where SVM can set up a separating hyper plane and linearly classify samples. Via the "kernel trick", it does not need to represent the feature space explicitly, simply by defining a kernel function $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, which substitutes of the dot product in the feature space.

The decision function of SVM based on the single kernel for binary classification is an $\boldsymbol{\alpha}$ weighted linear combination of kernels with a bias b as follows:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i \, y_i K(x_i, x) + b) \tag{1}$$

3.2 The Traditional Multiple Kernel Method

Multiple kernel methods extend the single kernel further to a convex linear combination of K basic kernels:

$$K(x_i, x) = \sum_{k=1}^{K} \beta_k K_k(x_i, x) \;\text{ with }\; \sum_{k=1}^{K} \beta_k = 1 \text{ and } \forall k : \beta_k \geq 0 .$$

Correspondingly, the decision function of the multiple kernels learning (MKL) problem can be formulated as [20]:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i \sum_{k=1}^{K} \beta_k K_k(x_i, x) + b) \tag{2}$$

where the optimized coefficients $\boldsymbol{\alpha}$ can be used to stand for the importance of every training sample, and the kernel weight $\boldsymbol{\beta}$ measures the importance of different basic kernels for the discrimination.

3.3 Per-Sample Based Multiple Kernel Method

In classical multiple kernel method, all training samples share the same kernel weight $\boldsymbol{\beta}$ within the same concept, only considering the importance of the corresponding basic kernel for learning such concept. However, treating training samples in a uniform manner neglects the personality of instance. In fact, the importance of different basic kernels relates to the training sample. Hence, we reformulate the problem formulated in Eqn.(2) by replace the basic kernel weight β_k of γ_{ik}:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i \sum_{k=1}^{K} \gamma_{ik} K_k(x_i, x) + b) \tag{3}$$

where γ_{ik} measures how the i-th training sample affects the discrimination in different basic kernel space. Compared with β_k in traditional multiple kernel method, γ_{ik} is no longer independent with the training sample. Accordingly, the number of the basic kernel weight rises from K to N×K.

Our goal is to optimize the coefficients $\boldsymbol{\alpha}$ and the kernel weight $\boldsymbol{\gamma}$ for constructing the decision function $f(x)$ that can capture both the commonness of visual concept and diversity of the concept instance.

It is essential to note that, each basic kernel $K_k(x_i, x_j)$ could uses a distinct set of features of x_i and x_j, representing the similarity of the two images on a certain aspect. Furthermore, different basic kernels can use different kernel forms, which can simply be classical kernels (such as Gaussian or polynomial kernels) with different parameters, or defined specially considering the specialty associated with image. Details on the basic kernel forms utilized in this paper will be introduced in Sec. 5.1.

4 Learning the Classifier

We present in this section the mathematical formulation of our per-sample multiple kernel learning (PS-MKL) problem and the algorithm we proposed for optimizing the parameters in PS-MKL.

The PS-MKL Primal Problem

In the PS-MKL problem, the sample x_i is translated via a mapping $\{\phi_k(x) \mapsto \square^{D_k}\}_{k=1}^{K}$ from the input space into K feature spaces $(\phi_1(x_i), \cdots, \phi_K(x_i))$ where

D_k denotes the dimensionality of the k-th feature space. For each feature map there will be a separate weight vector $\mathbf{w_k}$. Here we consider linear combinations of the corresponding output functions:

$$f(x) = \sum_{k=1}^{K} \gamma_k^{'} \left\langle \mathbf{w_k}, \phi_k(x) \right\rangle + b \tag{4}$$

The mixing coefficients $\gamma_k^{'}$ should reflect the utility of the respective feature map for the classification task. Inspired by SVM learning process, training can be implemented as the following optimization problem which involves maximizing the margin between training examples of two classes and minimizing the classification error:

$$\begin{aligned} &\min_{\beta,w,b,\xi} \ \frac{1}{2}\sum_{k=1}^{K} \gamma_k^{'} \left\| \mathbf{w}_k \right\|^2 + C\sum_{i=1}^{N} \xi_i \\ &\text{s.t. } y_i(\sum_{k=1}^{K} \gamma_k^{'} \left\langle \mathbf{w}_k, \phi_k(x_i) \right\rangle + b) \geq 1-\xi_i \quad \xi_i \geq 0 \ \ \forall i; \ \ \gamma_k^{'} \geq 0 \ \ \sum_{k=1}^{K} \gamma_k^{'} = 1 \ \forall k \end{aligned} \tag{5}$$

where $\left\| \mathbf{w}_k \right\|^2$ is a regularization term which is inversely related to margin, $\sum_{i=1}^{N} \xi_i$ measures the total classification error.

The PS-MKL Dual Problem

Via introducing Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$ into the above inequalities constraint, formulating the Lagrangian dual function which satisfies the Karush-Kuhn-Tucker(KKT) conditions, and the addition of the kernel weights substitution, the former optimization problem further reduces to a convex program problem as follows:

$$\begin{aligned} &\max_{\gamma} \ \min_{\alpha} \ J(\cdot) \\ &J(\cdot) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\sum_{k=1}^{K} \gamma_{ik} K_k(x_i, x_j)) - \sum_{i=1}^{N} \alpha_i \\ &\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0 \quad 0 \leq \alpha_i \leq C \ \ \forall i; \ \ \gamma_{ik} \geq 0 \ \ \sum_{k=1}^{K} \gamma_{ik} = 1 \ \ \forall i \ \ \forall k \end{aligned} \tag{6}$$

This max-min problem is the PS-MKL dual problem which relative to the primal problem as Eqn.(5). $J(\cdot)$ is a multi-object function for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ .For fixed parameters $\boldsymbol{\gamma}$, minimization of $J(\cdot)$ over the sample weights $\boldsymbol{\alpha}$, which equals to minimize the global classification error and maximize the inter-class interval. When $\boldsymbol{\alpha}$ is fixed, maximizing $J(\cdot)$ over kernel weight $\boldsymbol{\gamma}$ means to maximize the global intra-class similarity and minimize the inter-class similarity simultaneously.

Solving this Max-Min problem is a typical saddle point problem. Details of the solving procedure will be presented in the next part.

An Efficient Learning Algorithm

We adopt a two-stage alternant optimization approach, which is the traditional MKL route for parameter learning. Optimizing the sample coefficients α is to estimate which training samples are representative for learning this visual concept and the corresponding

weightiness for discrimination. Fixing the kernel weights $\boldsymbol{\gamma}$, the sample weight $\boldsymbol{\alpha}$ can be estimated by minimizing $J(\cdot)$ under the constraint $\sum_{i=1}^{N} \alpha_i y_i = 0$ and $\forall i : 0 \le \alpha_i \le C$. It can be reduced to the SVM dual quadratic program (QP) problem with a mixed kernel $K(x_i, x_j) = \sum_{k=1}^{K} \gamma_{ik} K_k(x_i, x_j)$. Consequently, minimizing $J(\cdot)$ over $\boldsymbol{\alpha}$ can be easily implemented as there already exists several efficient solvers for the corresponding QP.

Optimize the basic kernel coefficient γ_{ik} is in the interest of understanding which regulation on $\boldsymbol{\alpha}$ promotes the sparsity of kernel weights. Hence if one would be able to obtain an accurate classification by a sparse kernel weights, then one can quite easily interpret the resulting decision function.

To conveniently optimize the model parameters γ_{ik}, the objective function concerning γ_{ik} can be rewritten as:

$$J(\cdot) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} S_{ik}(\alpha) - \sum_{i=1}^{N} \alpha_i \quad , \tag{7}$$

where

$$S_{ik}(\alpha) = \frac{1}{2} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K_k(\mathbf{x}_i, \mathbf{x}_j)$$

Then the optimization of $J(\cdot)$ over γ_{ik} turns to:

$$\begin{aligned} &\max \ \theta \\ &\text{s.t.} \ \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} S_{ik}(\alpha) - \sum_{i=1}^{N} \alpha_i \ge \theta; \gamma_{ik} \ge 0 \ \sum_{k=1}^{K} \gamma_{ik} = 1 \quad \forall i \quad \forall k \\ &\quad \text{for all } \alpha \text{ with } \sum_{i=1}^{N} \alpha_i y_i = 0 \quad 0 \le \alpha_i \le C \end{aligned} \tag{9}$$

For the fixed optimal $\boldsymbol{\alpha}$, the optimization of $J(\cdot)$ over $\boldsymbol{\gamma}$ is a linear program (LP) because θ and $\boldsymbol{\gamma}$ are only linearly constrained. Nevertheless, the constraint on θ has to hold for every compatible α resulting in infinite constraints. In order to solve this so-called semi-infinite linear program (SILP) problem, a column generation strategy is employed as follows: Solving the former QP from a fixed $\boldsymbol{\gamma}$ produces a special $\boldsymbol{\alpha}$, which then increase a constraint on θ. We add the new constraints iteratively in this way and solve the LP with all gained constraints. This procedure has been proven to be converged in [20].

5 Experiments

In this section, we show the studies on Caltech 101 and Wikipedia MM data set. Compared with Caltech101, Wikipedia MM dataset is more close to the web images where instants of concept have more variation in appearances. Our primary goal is to investigate the effectiveness of the proposed multiple kernels method on open data set and practical retrieval data set and how does the performance change with the difficulty of the problem.

For Caltech 101, we select N images from each class for 1-vs-all training, e.g. N = 10, 20, 30. The remaining images are used for testing. In Wikipedia image data set, about 150,000 images are clawed from Wikipedia with diverse topics. These images are associated with unstructured and noisy textual annotations in English. Fifteen teams engaged in Image CLEF 2008 Wikipedia MM task [6] are required to submit at most 1000 related image in the dataset for each of the 75 predefined topics. Among these 75 topics, 10 topics, which have more than 100 positive samples over the results submitted by all fifteen participants, are picked out for our experiments. About 100 samples including positive and negative image are used for each topic's one-vs-all training. The remaining images in each topic are used for testing. It is worthy to note that images submitted by fifteen participants are searched via visual or text retrieval and highly related with the topics.

5.1 Features and Kernels

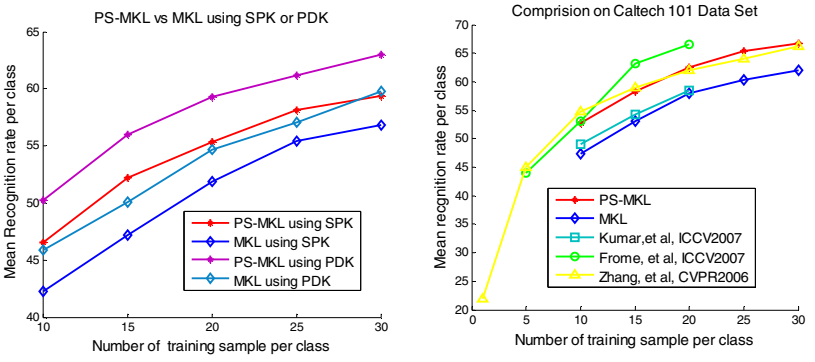
SIFT and Dense Color-SIFT are employed to characterize the local region of the image. For SIFT, SIFT descriptor is computed over the interest regions extracted by Different-of Gaussian (DoG) detector, forming a 72(3×3×8) dimensional feature for each interest region. For Dense Color-SIFT, SIFT descriptor is computed on RGB 3-channels over 16×16 pixel patch with spacing of 8 pixels, forming a 3×72 dimensional feature for each patch. SIFT normalization is skipped when the gradient magnitude of the patch is too weak. We use k-means algorithm to quantize the extracted descriptors from the training images to obtain codebooks with the size of k(400) for SIFT and Dense Color-SIFT respectively.

We use our own implementation of two types of latest kernels for object recognition, Spatial Pyramid Kernels (SPK) [18] and Proximity Distribution Kernels (PDK) [19] as basic kernels. For SPK, image is divided into $2^l \times 2^l$ cells and features from the spatially corresponding cells are matched across any two images. The resulting kernel is a weighted combination of histogram intersections from the coarse cells to the fine cells. Four levels pyramid is used with grid sizes 8×8, 4×4, 2×2 and 1×1 respectively. For PDK, local feature distributions of the r-nearest neighbors are matched across two images. The resulting kernel is a combination of multiple scale of the local distribution, e.g. r = 1,…, k. k is set to (8, 16, 32, 64) from finest to coarsest neighborhood in our implementation.

5.2 Experiment Results

Caltech 101: Our first experiment is carried on the Caltech101dataset. As mentioned above, 4 Spatial Pyramid Kernels (SPK) and 4 Proximity Distribution Kernels (PDK) over two types of local features are utilized as basic kernels, forming totally 16 basic kernels. In Fig. 4(a), we show the mean recognition rate of the MKL and PS-MKL over two types of basic kernels. Generally, PDK based methods achieve better results than the SPK based methods. Notice that PS-MKL outperforms MKL in any types of basic kernels.

In Fig.4 (b), the performances of MKL and PS-MKL over all basic kernels are compared with other works in [8,21,22]. When training sample is less than 20, performance of MKL is slightly lower than that in [8], which adopt a similar multiple kernels methods. One possible reason is different implementation for feature quantization.



(a) (b)

Fig. 4. Recognition results on Caltech 101 dataset

When MKL is fed with more training samples, disparity trends to be smaller. Overall, PS-MKL outperforms other traditional MKL methods. When the number of training samples for each object class reaches 30, PS-MKL obtains a performance of 66.7%, which is competitive with the latest reported results.

To summarize, PS-MKL has outperformed the existing MKL methods on the Caltech 101 dataset. More significantly, our approach can easily adopt and automatically weight the new basic kernels at sample level.

Wikipedia MM: For the Wikipedia MM dataset, we report the results using MKL and PS-MKL over the all basic kernels of SPK and PDK as before. With the regard of topic detection over thousands of image, we use F-score to measure the accuracy and effectiveness of concept learning. F-measure of each topic and its corresponding precision and recall are list in table 1 for MKL and PS-MKL.

Table 1. Recognition results on Wikipedia MM dataset

Topic	MKL			PS-MKL		
	Precision	**Record**	**F-score.**	**Precision**	**Record**	**F-score.**
Bridges	56.47%	76.42%	64.95%	54.88%	82.93%	**66.05%**
Cities by night	81.53%	80.56%	**81.04%**	79.75%	80.56%	80.15%
Football stadium	68.74%	62.51%	65.47%	69.81%	62.50%	**65.94%**
Historic castle	79.49%	82.35%	80.90%	72.72%	91.44%	**81.91%**
House architecture	68.94%	77.81%	73.11%	74.05%	76.80%	**75.39%**
Hunting dog	69.52%	60.71%	64.82%	59.66%	79.46%	**68.15%**
Mountains & sky	80.7%	84.4%	82.51%	80.80%	85.48%	**83.07%**
Military Aircraft	75.72%	58.82%	66.21%	72.52%	70.58%	**71.54%**
Race car	59.58%	86.67%	70.61%	64.62%	80.00%	**71.49%**
Star galaxy	84.99%	79.17%	**81.97%**	79.21%	79.16%	79.19%

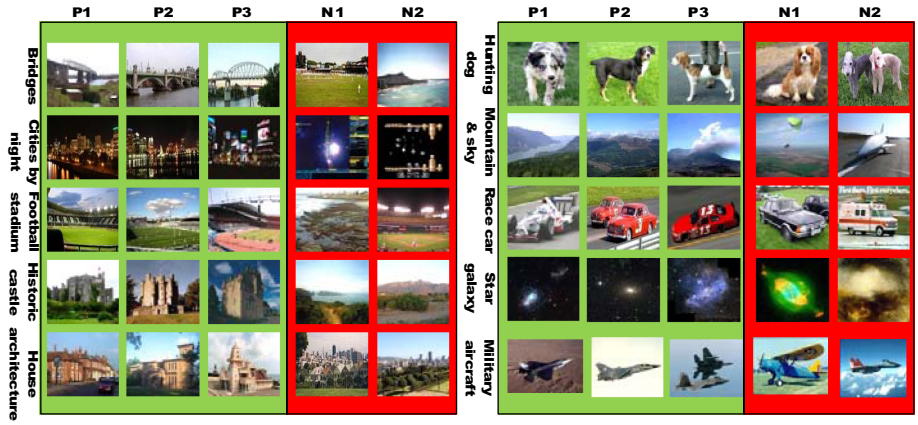


Fig. 5. Top three of the positive samples and top two of the negative samples in the recognized results on Wikipedia MM dataset

As shown in table1, comparable results are obtained in topics of "cities by night" and "star galaxy". With our sample based kernel weighting algorithm, PS-MKL archives different levels of improvement over MKL on the other eight topics. An interest observation is that three topics, "bridges", "house architecture" and "hunting dog", with bigger intra-class variation receive obvious improvement. One explanation is that PS-MKL outperforms MKL in seizing the characteristics of the training samples without losing generally discriminative ability.

To illustrate the affects of the concept learning, we show in Fig. 5, for each topic the top related positive and negative images. Note that some negative and positive images not only share the similar appearances, but also have correlations on semantic level, e.g. hunting dog with pet dog, and race car with vehicle.

6 Conclusion and Future Works

In this paper, we propose a novel multiple kernels approach called PS-MKL method for visual concept learning. Different from the traditional MKL methods, where kernels are weighted uniformly on concept level, PS-MKL aims to capture the distinct discriminative capability for each training sample in different basic kernel spaces. The proposed approach simultaneously optimizes the weights of the basic kernels for each training sample and the associated classifier in a supervised learning manner. The optimal parameters can be solved alternatively with off-the-shelf SVM solvers and simplex LPs. As shown in the experimental results on Caltech101 and Wikipedia MM dataset, PS-MKL archives significant improvement over the traditional MKL methods and competitive results with the state-of-the-art approaches for visual concept learning. PS-MKL provides a scalable solution for both combining large numbers of kernels and learning visual concept from a small training set.

We will continue our future works in two directions. First, we will add more elaborately designed basic kernels not only on visual feature, but also other types such

as text feature, to improve the performance. Second, we will explore the faster learning algorithm to optimize the increasing number of model parameters.

Acknowledgments

This work is supported by grants from Chinese NSF under contract No. 60605020, National Basic Research Program (973) of China under contract No. 2009CB320906, and National Hi-Tech R&D Program (863) of China under contract No. 2006AA010105.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. PAMI 22(12), 1349–1380 (2000)
2. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: ICCV Workshop on Content-based Access of Image and Video Databases, Bombay, India, pp. 42–50 (1998)
3. Vogel, J., Schiele, B.: Natural Scene Retrieval Based on a Semantic Modeling Step. In: Proc. Int'l. Conf. Image and Video Retrieval (July 2004)
4. Sivic, J., Russell, B., Efros, A., Zisserman, A.: Discovering Objects and Their Location in Images. In: Proceedings of the IEEE ICCV 2005, pp. 370–377 (2005)
5. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning Object Categories from Google's Image Search. In: Proceedings of the Tenth ICCV 2005, vol. 2, pp. 1816–1823 (2005)
6. http://www.imageclef.org/2008/wikipedia
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In: Conference on Computer Vision and Pattern Recognition Workshop (2004)
8. Kumar, A., Sminc, C.: Support Kernel Machines for Object Recognition. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, October 14-21, 2007, pp. 1–8 (2007)
9. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: Proc. Computer Vision and Pattern Recognition (2005)
10. Fei-Fei, L., Fergus, R., Perona, P.: One-Shot learning of object categories. IEEE Trans. PAMI 28(4), 594–611 (2006)
11. Jia, L., Fei-Fei, L.: What, where and who? Classifying event by scene and object recognition. In: ICCV (2007)
12. Ng, A., Jordan, M.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: Advances in NIPS, vol. 12 (2002)
13. Malisiewicz, T., Efros, A.A.: Recognition by Association via Learning Per-exemplar Distances. In: CVPR (June 2008)
14. Torralba, A., Fergus, R., Freeman, W.T.: Tiny images.Technical Report MIT-CSAIL-TR-2007-024, MIT CSAIL (2007)
15. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
16. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: ICCV, October 17-21, 2005, vol. 2, pp. 1458–1465 (2005)

17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
18. Ling, H., Soatto, S.: Proximity Distribution Kernels for Geometric Context in Category Recognition. In: ICCV, October 14-21, 2007, pp. 1–8 (2007)
19. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the SMO algorithm. In: NIPS (2004)
20. Sonnenburg, S., Raetsch, G., Schaefer, C., Scholkopf, B.: Large scale multiple kernel learning. Journal of Machine Learning Research, 1531–1565 (2006)
21. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification. In: ICCV 2007, pp. 1–8 (2007)
22. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In: CVPR. pp. 2126–2136 (2006)