S³MKL: Scalable Semi-Supervised Multiple Kernel Learning for Real-World Image Applications

Shuhui Wang, *Student Member, IEEE*, Qingming Huang, *Senior Member, IEEE*, Shuqiang Jiang, *Senior Member, IEEE*, and Qi Tian, *Senior Member, IEEE*

Abstract—We study the visual learning models that could work efficiently with little ground-truth annotation and a mass of noisy unlabeled data for large scale Web image applications, following the subroutine of semi-supervised learning (SSL) that has been deeply investigated in various visual classification tasks. However, most previous SSL approaches are not able to incorporate multiple descriptions for enhancing the model capacity. Furthermore, sample selection on unlabeled data was not advocated in previous studies, which may lead to unpredictable risk brought by real-world noisy data corpse. We propose a learning strategy for solving these two problems. As a core contribution, we propose a scalable semi-supervised multiple kernel learning method (S³MKL) to deal with the first problem. The aim is to minimize an overall objective function composed of log-likelihood empirical loss, conditional expectation consensus (CEC) on the unlabeled data and group LASSO regularization on model coefficients. We further adapt CEC into a group-wise formulation so as to better deal with the intrinsic visual property of real-world images. We propose a fast block coordinate gradient descent method with several acceleration techniques for model solution. Compared with previous approaches, our model better makes use of large scale unlabeled images with multiple feature representation with lower time complexity. Moreover, to address the issue of reducing the risk of using unlabeled data, we design a multiple kernel hashing scheme to identify the "informative" and "compact" unlabeled training data subset. Comprehensive experiments are conducted and the results show that the proposed learning framework provides promising power for real-world image applications, such as image categorization and personalized Web image re-ranking with very little user interaction.

Index Terms—Image categorization, multiple kernel learning, multiple kernel locality sensitive hashing, personalized image re-ranking, semi-supervised learning.

I. INTRODUCTION

DUE to the popularity of the digital camera and the ubiquitousness of the Internet, digital photos and videos are produced, uploaded and exchanged online every day, leading to the explosive growth of Web multimedia. For such a huge database with massive un-annotated content, effective methods for Web media data mining are indispensable. One major solution for this urgent need is to train a set of automatic annotators on well-labeled images, and annotate the new images with one or more visual concepts. Then users may easily find those images that contain the visual concepts by querying with text. For the past decade, numerous literatures have been devoted to this area [5], [9], [10], [22], [28], [32], [36], [39], [43], [45], [46].

Although promising results have been reported on various image datasets, the visual classification is still far from being perfect for successful real-world application because previous approaches can only work on dataset with thousands of well-chosen images. To enhance the model capacity, the most direct way is to obtain sufficient well-labeled training examples. Although a lot of efforts have been devoted to building large scale ground truth database, such as ImageNet [7], the growth of labeled data would never catch up with the growth of new un-annotated images on the Web. Therefore, semi-supervised learning (SSL) [47], which is extended from supervised learning and provides the potential to take advantage of unlabeled data for model enhancement, would be a better choice for real-world applications. However, previous SSL approaches can only work on small scale and clean datasets due to many reasons. In this paper, we address several issues that hinder SSL approaches from real-world application.

The first issue is closely related to some intrinsic weaknesses of SSL. Single image feature lacks sufficient discrimination power under uncontrolled appearance changes in real world such as light, occlusion and various viewing angles, which has been a long unsolved difficulty for learning a good model. A practical solution is to combine the discrimination powers of heterogeneous but complementary features and their corresponding similarity measures. To this end, multiple kernel learning (MKL) [2], [18], [23], [29] and its variants [1], [5], [11], [36], [45] have been applied to image classification. Unlike previous single kernel approach [37], MKL is capable of approximating the optimal similarity measures by optimizing the weights of the linear combination of a set of kernels, and minimizing the objective function simultaneously.

However, studies have rarely been done, except Tsuda *et al.* [35], on how MKL employs unlabeled data to improve the prediction power of the model because of the difficulty of defining

a proper loss on the unlabeled data with multiple features/kernels. Being a direct extension from graph based SSL [48] and the original uniform MKL [2], [18], [23], [29], the method in [35] is transductive, which could not flexibly predict an unseen sample. Also, it is not suitable for large scale image classification since both expensive matrix inverse computations and huge memory are required.

To overcome the weaknesses of previous SSL approaches, we propose S^3MKL, an inductive and scalable semi-supervised multiple kernel learning method with *conditional expectation consensus (CEC)*, a carefully designed conditional expectation based regularization on unlabeled data. Inheriting the scalability from the expectation regularization [19], [20], this new regularization framework is suitable for large scale SSL, which penalizes the expectation inconsistence between labeled data and unlabeled data with respect to different kernels. Moreover, we are inspired by the studies of adapting existing machine learning approach to visual learning problems such as [24] and [45]. Both Yang *et al.* [45] and Saffari *et al.* [24] proposed some "group sensitive" measurements among distinct groups of image subset in the objective function rather than uniform measure among the entire dataset. Such an intermediate group structure has been proved to be more effective than the uniform structure to overcome the difficulty brought by the notorious "intra-class diversity" and "inter-class correlation" in visual learning domain. To cope with the complicated real-world visual distribution consistently brought by "intra-class diversity" and "inter-class correlation", we further extend the *conditional expectation consensus* on training images that have been divided into several groups with visually similar images, which ensures the model expectation on unlabeled data better fits to the reference expectation on labeled data.

Furthermore, the discriminate model used by S^3MKL is kernel logistic regression which incorporates multiple kernels in a sample specific way [5], [11], [45], providing more precise feature/kernel fusion capability. For the regularization of the model coefficients, we adopt group LASSO to ensure the sparsity over group level, which has been proved to be more effective than l_1-norm and l_2-norm if the group structure exists among the model coefficients [1], [14], [21].

To effectively train the model, we propose a fast and robust optimization method based on block coordinate gradient descent method (BCGD) [34] which is suitable for minimizing objective functions with l_1-norm and group LASSO. By employing two acceleration techniques, namely, the heuristic re-allocation and random permutation, our optimization method converges 6 times faster than the original BCGD [34].

The second limitation preventing previous approaches from real-world applications is that the model takes risk to be corrupted by real-world noisy information if no training data selection system is constructed. Previous approaches have proved that conducting labeled sample selection can help reduce the training time and enhance the model robustness [46]. Besides these benefits, we also stress in this paper that labeled data selection can be helpful for certain real-world application such as personalized image re-ranking that needs some user annotation as the input and a personalized re-ranking model should be quickly learned based on the data which encodes the preference of individual user.

However, what kind of unlabeled data can be helpful for model enhancement have been less addressed in previous studies, where most of them assumed implicitly that the distribution of the unlabeled data is similar to the labeled data, or there are definitely some positive examples existing among the unlabeled data corpus. A theoretical analysis on how unlabeled data will help has been provided in [27]. Nevertheless, most of the assumptions and analysis are too strict to be generalized into Web application where noisy images usually pervade. In this paper, our intuition comes from two common senses. Firstly, a model could not accurately predict an unlabeled image that is very visually different from the labeled training samples, especially the positive labeled samples. Secondly, Web images containing much irrelevant or noisy content (e.g., logo images and advertisement images) are likely to be a risk of degradation for semantic object detectors. When these two situations take in place, the model will be likely to degrade if we do not impose certain control. Therefore, we argue that sample selection is indispensable not only for labeled training data but also for unlabeled training data.

To efficiently conduct training sample selection, we construct an approximated nearest neighbor search system based on the kernel version of LSH (KLSH) [16]. To combine the discrimination power of different features/kernels, we build a multiple kernel LSH (MKLSH) based on a set of KLSH functions with respect to several kernels. The advantage over a single kernel hashing is that it can reduce the bias caused by single metric by combining multiple visual properties.

In summary, the main contributions of our work are:

- S^3MKL, a scalable semi-supervised multiple kernel learning method for classifier training with labeled and unlabeled data based on multiple kernel representation.
- A much faster coordinate gradient descent optimization method for model training which is about 6 times faster than the original approaches.
- MKLSH, an approximated nearest neighbor search system based on multiple KLSH which conducts labeled and unlabeled sample selection for real-world image application.
- Solutions using S^3MKL and MKLSH for real-world image categorization and personalized Web image re-ranking.

Section II reviews related works. In Section III, we introduce S^3MKL and the optimization procedure. In Section IV, the detail of MKLSH is discussed. In Section V, we present how to combine S^3MKL and MKLSH for image classification and personalized image re-ranking. Section VI presents the experiments. Section VII provides conclusion and future work.

II. RELATED WORKS

Our work is mainly related to four topics, image annotation and object recognition, learning from real-world images, semi-supervised learning, and learning with multiple features/kernels. Due to the limited space, we only brief some recent and representative works on these four aspects.

A. Image Annotation and Object Recognition

Image annotation has been a hot issue in computer vision and multimedia communities. A well-studied paradigm, as we

call annotation by learning, is to train a set of automatic annotators based on a collection of labeled image using learning method such as SVM [9], [10], [28] and adaboost [43], [32]. To better facilitate the existing method for real-world image understanding, a lot of research has been conducted to address the problem from different point of views. Qi *et al.* [22] proposed a large margin learning approach for multi-label classification which encodes the co-occurrence among different semantic labels into kernel computation. Saffari *et al.* [24] and Tang *et al.* [30] developed two different semi-supervised image categorization methods. Torralba *et al.* [32] and Yan *et al.* [43] studied the feature and model sharing phenomenon among image categories and developed effective and scalable learning models for multi-class categorization. Torralba *et al.* [31] showed that a naïve k-NN classifier using carefully designed similarity measure on 80 million images collected from the Web will achieve good results. Along this routine, a nearest neighbor based approach was proposed by [46], in which every test sample is predicted by an SVM classifier trained on its nearest k samples. It inherits advantages from both k-NN and SVM. While their aim is to find the training samples that are informative to the test samples, we apply nearest neighbor search to find some informative unlabeled data.

Recently, multiple kernel learning [2], [18], [23], [29] and its variants [1], [5], [11], [36], [45] have been successfully applied in visual recognition domain. When working in visual analysis domain, MKL provides a way to reach a better tradeoff between capturing the diversity while keeping the invariance of semantic categories by combining the discriminate power of each feature/kernel. The success of MKL also inspired researchers such as Duan *et al.* [8] to extend current transfer learning model so that multiple kernels can be incorporated for seeking better generalization power.

B. Learning From Real-World Images

The study of visual learning usually suffers from the shortage of ground truth data. To alleviate this problem, there are mainly two sub-routines. Firstly, many studies have been devoted to building large scale database with structure semantic hierarchies [7], [31] with the cooperative work of human and artificial intelligence [7], [31], [33], [44].

The second sub-routine takes the following statement as the common knowledge, i.e., the growth of annotated images cannot catch up with the growth of new unlabeled images. Based on this consideration, to efficiently exploit the abundant unlabeled image resources, semi-supervised learning [24], [30] is studied in the context of image application. Recently, there are growing interests in solving annotation and object recognition by exploiting the nearly infinite Web repository [8], [31], [41], [42]. The advantage of this method is that it can deal with hundreds and thousands of image categories by utilizing the abundant tagging information provided by Web users. These methods are flexible in processing popular queries such as images of human or pets. However, the models will degrade when the returned Web images contains fewer true tags and more noise, especially when considering the fact that the user tags usually include personal preferences. In this case, we argue that annotation by learning still plays an important role.

For visual learning using large scale data corpus, a critical problem is to quickly identify a subset of semantically relevant images since learning with the whole data is prohibitive. To this end, approximated nearest neighbor search such as locality sensitive hashing [6] can be an efficient solution. Since the canonical LSH can only hash data samples in the original Euclidean feature space, the hit rate of true nearest neighbors will be degraded when processing image data. This difficulty inspired hashing techniques with kernel representation [16] and learned metrics [17]. Our study is inspired by [16] as we extend the single kernel hashing technique to multiple-kernel hashing, which shows better performance for various genres of image data retrieval.

C. Semi-Supervised Learning

Instead of only minimizing the empirical risk or structural risk, semi-supervised learning also minimizes the risk defined on unlabeled data based on some assumptions, such as: 1) manifold assumption [25], [26], [48], which assumes that similar data would be more likely to have the same label; 2) the max-margin criterion [3], [15], [24], which prevents the classification boundary located on higher density areas, and maximizes the margin defined on both labeled and unlabeled data; 3) minimum entropy [12], in which the entropy of some conditional distribution on unlabeled data is minimized. To apply SSL for image annotation, in [30], the graph-based SSL [47], [48] is modified to incorporate the local density difference of samples, and achieves promising results on video concept annotation task. In [24], a semi-supervised multi-class boosting method is proposed, which is based on cluster assumption and expectation regularization [19], [20] is applied for regularization on unlabeled data, and the defined "*margin*" on both labeled and unlabeled data is maximized.

A relevant SSL of incorporating multiple features is co-training [4] where two classifiers boost each other by introducing new reliably predicted unlabeled samples to each other. This knowledge is then promoted in multi-view learning that multiple hypotheses are trained from different views of the same labeled data, and are required to make consistent predictions on any given unlabeled instance [25]. Multi-view kernel approach [26], [49] was proposed based on [25]. They can be easily distinguished from S^{3}MKL since the former trains several classifiers simultaneously, and ensembles classifiers in late fusion style [28], while S^{3}MKL exploits the discriminative power provided by multiple kernels in a single classifier.

The most related work with our study could be found in [35], where multiple feature networks are combined with graph based SSL. However, it is not capable of real-world image retrieval, because: 1) it is transductive and not flexible to predict unseen samples; 2) it requires expensive inverse matrix operation for objective function minimization; 3) uniform weight is assigned to the graph *Laplacian* of each kernel among all the samples, which have more restricted fusion capability than sample dependent kernel weight assignment [5], [11], [45].

D. Employing Multiple Features and Kernels

A lot of studies have been conducted on efficiently combining different features. A simple scheme is feature concatena-

tion [28], but feature sparseness and "curse of dimensionality" will be introduced. Another solution is the decision fusion [28] of multiple classifiers, but it just obtains a more stable decision based on the individual decision output.

As one of the most promising feature fusion methods, multiple kernel learning [2], [18], [23], [29] combines multiple features by linear kernel combination. It avoids "curse of dimensionality" incurred by feature concatenation. Specifically, an information-theoretic MKL with l_1-norm kernel weight was proposed by [50], [51], where the conditional entropy loss on the labeled data is minimized. MKL has also been used for object recognition [23], [36], and good results have been reported on many challenging datasets.

Early studies on MKL usually impose uniform kernel weighted combination to each sample. In more recent studies, kernel weight is assigned differently to each sample or each group of samples [11], [45]. Compared with the uniform kernel weight setting, the local weight approaches perform more favorable because the local kernel weight provides more deliberate fusion capability. The disadvantage of studies in [11] and [45] is that no global convergence is guaranteed. A new localized MKL is proposed by [5], [38], where group LASSO is applied directly on regularizing the local kernel coefficients, instead of fitting a gating function for the calculation of the coefficients as in [11], [45]. By optimization with BCGD [34], the model is guaranteed to converge to global optimal solution.

III. S³MKL

A. Definition and Objective Function

We are given N training samples of two classes: (x_i, y_i, o_i), $y_i \in \{0, 1\}$, $o_i \in \{0, 1\}$, $1 \le i \le N$, where $o_i = 1$ denotes labeled data and $o_i = 0$ denotes unlabeled data, and y_i denotes the label of the ith labeled training sample. We use $L = \{(x_i, y_i)|o_i = 1\}$ to denote the set of labeled data, and $U = \{(x_i)|o_i = 1\}$ for unlabeled data in the rest of the paper. We calculate M kernels for both labeled data and unlabeled data, so we denote each kernel by m, where $m = 1, \ldots, M$. The discriminative function $\mathbf{f}$ in our learning framework is a kernel logistic regression model integrating M different kernels. It is formulated as

$$\mathbf{f}(x) = \alpha_0 + \sum_{i=1}^{|L|} \boldsymbol{\alpha}_i \mathbf{K}_i(x) \tag{1}$$

where $\boldsymbol{\alpha}_i = [\alpha_i^1, \ldots, \alpha_i^M]$ is the unknown kernel logistic regression parameter of the ith sample and $\mathbf{K}_i(x) = [K_1(x_i, x), \ldots, K_M(x_i, x)]^T$ denotes the similarity measures between x_i and x with respect to M different kernels; α_0 is the bias term. Under the binary classification scenario, the probability of sample x belonging to the positive class $(+1)$ is calculated as: $P(1/x) = (1 + e^{-\mathbf{f}(x)})^{-1}$. It is very easy to extend this model to the multi-class version, but in this paper, we only demonstrate how our method works in binary classification. For a multi-class classification task, we adopt one-versus-all scheme. The predicted label is the class corresponding to the largest class probability.

We minimize the following objective function with respect to the classification function $\mathbf{f}$:

$$\min_{\mathbf{f}} Q(\mathbf{f}) = \min_{\mathbf{f}} Z(\mathbf{f}) + \gamma \Theta(\mathbf{f}) + \lambda \Omega(\mathbf{f}) \tag{2}$$

where $Z(\mathbf{f})$ denotes the weighted negative log-likelihood loss on labeled data; $\Omega(\mathbf{f})$ denotes group LASSO [1], [5], [21] penalty on $\mathbf{f}$, which ensures the model sparsity on group level, and $\Theta(\mathbf{f})$ denotes the *CEC* penalty on unlabeled data, which regularizes the behavior of different kernels on unlabeled data by minimizing the difference of the conditional expectation on unlabeled data and the given reference expectation. γ and λ are the weights of Θ and Ω, respectively. It is worth noting that when $\gamma = 0$, the objective function is identical to [5] and becomes a supervised learning method.

The loss on the labeled data is defined as the weighted negative log-likelihood as

$$Z(\mathbf{f}) = \sum_{x_j \in L} b_j \log(1 + \exp(\mathbf{f}(x_j))) - \sum_{x_j \in L} b_j y_j \mathbf{f}(x_j) \tag{3}$$

where b_j denotes the weight of the jth sample. b_j is introduced to avoid the class imbalance problem. In this paper, we set $b_j = 2$ and $b_j = 1$ for positive and negative data, respectively.

B. Group LASSO Regularization

Previous studies have shown that l_1 regularization will lead to the sparse solution of the model parameter [14]. However, in practical situation, the group structures usually exist on the model vector space, so that parameters in the same group tend to be zeros or non-zeros simultaneously. In this case, group LASSO [1], [5], [14], [21], which ensures the model is sparse on group level is more preferred. Huang *et al.* [14] studied the benefit of group LASSO, and provided a convincing theoretical justification for using group sparse regularization when the underlying group structure is consistent with the data. In our method, we define the "group" as all the kernel coefficients corresponding to each labeled sample, and the regularization term could be written as

$$\Omega(\mathbf{f}) = \sum_{i=1}^{|L|} ||\boldsymbol{\alpha}_i||_2. \tag{4}$$

The group LASSO regularization in (4) prevents those outlier samples to be chosen into the model by setting all the kernel coefficients of the sample to zero. This function leads to the sparsity of the kernel coefficients $\boldsymbol{\alpha}_i$ on group level. Comparison with SVM would be very interesting. Although the inherent mechanisms are quite different, for both methods only a small portion of samples will contribute to the final model.

In fact, larger λ will lead to sparser model coefficients and under-fitting, while smaller λ leads to denser model coefficients and over-fitting. In this paper, we found that $\lambda = 0.01$–0.05 is a reasonable choice to guarantee the performance and robustness. The setting of λ need not be precisely tuned for every dataset. After a preliminary validation process, for all the experiments in this paper, we set $\lambda = 0.02$.

C. Conditional Expectation Consensus

We propose *CEC* based on the idea from expectation regularization [19], [20], which is a scalable regularization framework for semi-supervised learning method, especially for exponential family parametric models. It augments the traditional conditional label-likelihood objective function with an additional term that encourages model predictions on unlabeled data to match certain reference expectations. This regularization is suitable for solving real-world classification problem. For example, the ratio of "*car*" images in a large set of Web images could be easily obtained from the Web even we do not know exactly whether a certain image contains a car. We could also obtain the knowledge of how likely an image would contain a car if the word "*car*" appears in the surrounding text. The former situation is known as the "label priors", and the latter is an example of "feature labeling" [19]. Both of them could be used as the reference expectation in the expectation regularization framework. The expectation regularization could be seen as a generalized version of entropy regularization [12]. However, there is no such function in expectation regularization for modeling the behavior of coefficients of different kernels.

To adapt expectation regularization for multiple kernel learning, we define the marginal kernel density (MKD) for each labeled data and unlabeled data as

$$g_m^\pi(x) = \left\langle \overline{\phi}_m^\pi, \psi_m(x) \right\rangle \qquad (5)$$

where $\psi_m(x) = [K_m(x_1, x), \ldots, K_m(x_L, x)]^T$ denotes the kernel values between sample x with all the labeled training data on the mth kernel channel. $\overline{\phi}_m^\pi$ is a non-negative random vector with each element independently generated from a distribution such as the uniform distribution $U(0, 1)$. To avoid model degradation brought by improper random projection, we normalize each $\overline{\phi}_m^\pi$ to ensure $||\overline{\phi}_m^\pi||_2 = 1$. The MKD measures the kernel response for sample x on the mth kernel channel. To reduce the projecting bias brought by using only one random vector, we adopt the criterion of projecting Π times for each kernel and each sample, and we denote the index of random projection as π. In our paper, we set $\Pi = 5$ for all the experiments, which guarantee good results. We apply expectation regularization on MKD. The reference expectation q_m^π and the model expectation p_m^π are calculated as

$$q_m^\pi(y|g_m^\pi(x)) = \frac{1}{G_m^\pi} \sum_{x \in L} y(x) g_m^\pi(x),$$

$$G_m^\pi = \sum_{x \in L} g_m^\pi(x)$$

$$p_m^\pi(\mathbf{f}(x)|g_m^\pi(x)) = \frac{1}{\overline{G}_m^\pi} \sum_{x \in U} P(\mathbf{f}(x)) g_m^\pi(x),$$

$$\overline{G}_m^\pi = \sum_{x \in U} g_m^\pi(x) \qquad (6)$$

where $P(\mathbf{f}(x)) = (1 + e^{-\mathbf{f}(x)})^{-1}$. The proposed *CEC* regularization is formulated as

$$\Theta(\mathbf{f}) = \frac{1}{\Pi} \sum_{\pi=1}^{\Pi} \sum_{m=1}^{M} D(q_m^\pi(y|g_m^\pi(x)) || p_m^\pi(\mathbf{f}(x)|g_m^\pi(x); \boldsymbol{\alpha})) \qquad (7)$$

where D measures the Kullback-Leibler divergence of q_m^π and p_m^π. The parameter γ determines the relevant importance of the penalty. We find that the performance of the model is not very sensitive to the value of γ, so it does not need careful tuning for each dataset. This observation is also consistent with [19] and [20]. Unless specified in this paper, we empirically set $\gamma = 0.1 \cdot |L|$ after a preliminary validation process.

D. Group-Wise Conditional Expectation Consensus

For real-world image data, the phenomenon of intra-class variance and inter-class correlation are usually very prominent, so that images from different classes will overlap while images from same classes are usually far from each other in high dimensional space due to various reasons such as background clutter and viewing angles. Therefore, sometimes it will be hard for the model expectation p_m^π to be regularized towards a single p_m^π when facing with complicated data distribution.

A good solution for dealing with this complicated situation is to divide the training image dataset into several groups, where images within a group share some common visual properties. Similar ideas can also be found in previous studies such as [24] and [45]. Therefore, to better facilitate our method into real-world problems, we divide the labeled data and unlabeled data into R groups $\omega(1), \ldots, \omega(R)$ satisfying

$$L \cup U = \bigcup_{r=1}^{R} \omega(r),$$

$$\forall r_1, r_2, \omega(r_1) \cap \omega(r_2) = \varnothing, 1 \le r_1 \ne r_2 \le R \qquad (8)$$

and correspondingly, the reference expectation and the model expectation are calculated as

$$q_{m,r}^\pi = \frac{1}{G_{m,r}^\pi} \sum_{\substack{x \in L \\ x \in \omega(r)}} y(x) g_m^\pi(x),$$

$$G_{m,r}^\pi = \sum_{\substack{x \in L \\ x \in \omega(r)}} g_m^\pi(x)$$

$$p_{m,r}^\pi = \frac{1}{\overline{G}_{m,r}^\pi} \sum_{\substack{x \in U \\ x \in \omega(r)}} P(\mathbf{f}(x)) g_m^\pi(x),$$

$$\overline{G}_{m,r}^\pi = \sum_{\substack{x \in U \\ x \in \omega(r)}} g_m^\pi(x) \qquad (9)$$

we modify the *CEC* into group-wise version as

$$\Theta(\mathbf{f}) = \frac{1}{\Pi} \sum_{\pi=1}^{\Pi} \sum_{m=1}^{M} \sum_{r=1}^{R} D\left(q_{m,r}^\pi || p_{m,r}^\pi\right). \qquad (10)$$

The regularization in (10) is a generalized version of (7), and (7) can be treated as a special case where $R = 1$. In some situations, there might be no labeled data in some cluster, so the reference expectation with respect to this cluster could not be computed. We can use the reference expectation on the whole labeled data as calculated in (6) as the reference expectation for this cluster. The setting of cluster number R determines how precise the model expectation locally fits the reference expectation. However, if R is too large, the model will not only take risk

to be over-fitted, but also incur extra computational burden for the clustering process. Typically, by parameter tuning on validation datasets we find $R = 10$ should be a reasonable choice for both robustness and efficiency.

The most common solution to divide the training data into groups is to conduct clustering algorithm such as K-means on the original feature space [24], [45]. In Section IV, we will discuss a fast clustering process based on the MKLSH system.

E. Analysis on Conditional Expectation Consensus

We provide some analysis on how the regularization in Sections III-C and III-D works. We denote the feature representation in the implicit *Hilbert* Space defined by kernel K_m as $\Phi_m(x)$. Naturally, for any x, we have $||\Phi_m(x)||_2 \leq 1$. When training data x_i and x_j are within a ε-ball, they satisfy

$$||\Phi_m(x_i) - \Phi_m(x_j)||_2 \leq \varepsilon, \ \forall \ m \tag{11}$$

then we have the following proposition for showing the relation of the MKD among similar data:

Proposition 1: When training data x_i and x_j are within an ε-ball and when $\lim \varepsilon = 0$, the MKD between them are bounded by $\rho = \varepsilon \cdot |L|$ and $\lim \rho = |L| \cdot \lim \varepsilon = 0$.

Proof: According to *Cauchy-Schwarz* inequality, we have

$$|g_m^\pi(x_i) - g_m^\pi(x_j)| = \left| \left\langle \overline{\phi}_m^\pi, \psi_m(x_i) - \psi_m(x_j) \right\rangle \right|$$

$$= \left| \sum_{l=1}^{|L|} \phi_m^\pi(l) \cdot \left\langle \Phi_m(x_i) - \Phi_m(x_j), \Phi_m(x_l) \right\rangle \right|$$

$$\leq \left| \sum_{l=1}^{|L|} \phi_m^\pi(l) \cdot ||\Phi_m(x_i) - \Phi_m(x_j)||_2 \cdot ||\Phi_m(x_l)||_2 \right|$$

$$\leq \varepsilon \sum_{l=1}^{|L|} |\phi_m^\pi(l)| \cdot ||\Phi_m(x_l)||_2 \leq \varepsilon |L| \cdot ||\overline{\phi}_m^\pi||_2 \leq \varepsilon |L|. \tag{12}$$

As Proposition 1 shows, when training samples are visually similar, they have similar MKD, but not *vice versa*. MKD is a smooth density estimation that reveals the cluster distribution with respect to different kernel channels of the training data. Some examples of MKD are shown in Fig. 1, where hot color and dark color represent regions with high MKD and low MKD, respectively. Using different random projection for each kernel channel leads to slightly different MKD map in the feature space, so that different locality bias would be introduced into the model. However, by averaging all the locality bias, our method will gain smaller model bias and variance. Based on the property of MKD, we have the following remarks.

Remark 1: For S^3MKL, see (7), the unlabeled data nearby the labeled data will have more weights (higher MKD), so their prediction will be regularized to be determined by their nearby samples. For unlabeled data far away from the labeled data, the regularization tends to be weaker so as to avoid improper regularization on noisy information.

This property of S^3MKL guarantees that unlabeled samples with high density area will be correctly classified, which is similar in spirit with SSL methods based on *max-margin* criteria [3]
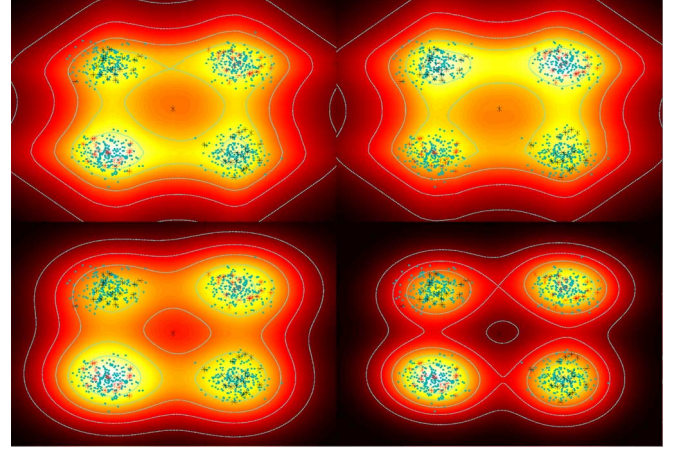

Fig. 1. Examples of marginal kernel density (best viewed in color). Dots in red, black, and green represent positive, negative, and unlabeled data, respectively. Each row represents one type of kernel with different random projection. The color of the region implies its density value, where hot color represents larger density.

or cluster assumption [12]. However, we find that S^3MKL does not make full use of unlabeled data because loose regularization is imposed on mid-distant unlabeled data. We have the following discussion to show the advantage of GS^3MKL on using mid-distant unlabeled data.

Remark 2: For GS^3MKL, see (10), the prediction of the unlabeled data nearby a cluster (group) of labeled data will be regularized to be determined by the labeled data in cluster. Therefore, better discriminative power on this cluster of unlabeled data can be achieved. For unlabeled data not belonging to any cluster with labeled data (no labeled data exists in their clusters), the conditional expectation on the unlabeled data will be regularized to the reference expectation calculated by all the labeled data.

From Remark 2 we see that some kind of "label prior" from the whole labeled data is imposed on those mid-distant unlabeled data. However, when the unlabeled data is too dissimilar with all the labeled data so that the prediction on them tends to be random, imposing such prior on them will gain no beneficial discriminative information but noise. In this case, unlabeled sample selection using MKLSH will be a necessary preprocessing for model enhancement.

One may argue with the effectiveness of GS^3MKL when the number of labeled data is usually limited, therefore introducing the group-wise cluster information will make the prior probability estimation on the labeled data less accurate without sufficient labeled data. However, the clusters do exist in many real-world visual datasets such as VOC [24], [45] and ImageNet. Considering the fact that the amount of unlabeled data is usually sufficient, in order to accurately estimate the group structure, we use both labeled and unlabeled data for cluster identification. By incorporating such local group structure, the learning model will be endowed with more precise classification boundary on those samples that are very hard to distinguish within a cluster. For those clusters with no labeled data, imposing a global conditional prior knowledge would not introduce any unstable local prior estimation from the labeled data.

F. Optimization

To effectively minimize $Q(\mathbf{f})$, we propose a new solution based on BCGD method by Tseng et al. [34]. BCGD was also employed in [5] and [21], solving group LASSO logistic regression in different contexts. Compared with the original BCGD, for learning problems on very unbalanced dataset, our method achieves better convergence rate. We introduce the details as follows.

Since the group LASSO regularization is not differentiable everywhere, Tseng et al. [34] suggest that a good way of minimizing this objective function is to decompose $Q(\mathbf{f})$ into a series of differentiable sub-problems with respect to $\boldsymbol{\alpha}_i$. Quadratic approximation and line search is combined to solve every sub-problem.

Firstly, the overall loss function is rewritten as

$$\min_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha}) = \min_{\boldsymbol{\alpha}} C(\boldsymbol{\alpha}) + \lambda \sum_{i=1}^{|L|} ||\boldsymbol{\alpha}_i||_2$$
$$C(\boldsymbol{\alpha}) = Z(\boldsymbol{\alpha}) + \gamma\Theta(\boldsymbol{\alpha}). \tag{13}$$

For each step, the loss function is firstly approximated by Taylor expansion:

$$Q(\boldsymbol{\alpha} + \mathbf{d}) \approx C(\boldsymbol{\alpha}) + \nabla C \cdot \mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathbf{H}\mathbf{d} + \lambda \sum_i ||\boldsymbol{\alpha}_i + \mathbf{d}_i||_2 \tag{14}$$

where $\mathbf{d}$ denotes the direction in which $\boldsymbol{\alpha}$ should be updated. $\mathbf{H}$ is a diagonal matrix approximating the Hessian of $C(\boldsymbol{\alpha})$ with the form

$$\mathbf{H} = diag(\mathbf{H}_1, \ldots, \mathbf{H}_{|L|})$$
$$\mathbf{H}_i = h_i\mathbf{I}, \ h_i = \max\left(diag(\nabla_{ii}^2 C), 10^{-3}\right). \tag{15}$$

Since $\mathbf{H}$ is diagonal and separable, we alternatively optimize the decomposed sub-problems:

$$Q_i = C(\boldsymbol{\alpha}) + \nabla_i C \cdot \mathbf{d}_i + \frac{1}{2}\mathbf{d}_i^T\mathbf{H}_i\mathbf{d}_i + \lambda||\boldsymbol{\alpha}_i + \mathbf{d}_i||_2. \tag{16}$$

We denote

$$\mathbf{k}_i(x_j) = [K_1(x_i, x_j), \ldots K_M(x_i, x_j)], i = 1, \ldots, |L|$$
$$s_{m,r}^\pi = \frac{1}{\overline{G}_{m,r}^\pi}\left(\frac{q_{m,r}^\pi}{p_{m,r}^\pi} - \frac{1 - q_{m,r}^\pi}{1 - p_{m,r}^\pi}\right)$$
$$v_{m,r}^\pi = \frac{1}{\overline{G}_m^\pi}\left(\frac{q_m^\pi}{(p_m^\pi)^2} + \frac{1 - q_m^\pi}{(1 - p_m^\pi)^2}\right) \tag{17}$$

then the gradient is calculated as

$$\nabla_i C = \nabla_i Z + \nabla_i \Theta$$
$$= \sum_{j=1}^N b_j(P_j(1) - y_j)\mathbf{k}_i(x_j)$$
$$+ \frac{\gamma}{\Pi}\sum_{\pi=1}^{\prod}\sum_{m=1}^M\sum_{r=1}^R s_{m,r}^\pi \sum_{x \in U, x \in \omega(r)} P_x(1)\mathbf{k}_i(x)g_m^\pi(x) \tag{18}$$

and the Hessian is calculated as

$$\nabla_{ii} C = \nabla_{ii} Z + \nabla_{ii}\Theta$$
$$= \sum_{j=1}^N b_j P_j(1) P_j(0)\mathbf{k}_i^2(x_j)$$
$$+ \frac{\gamma}{\Pi}\sum_{\pi=1}^{\prod}\sum_{m=1}^M\sum_{r=1}^R s_{m,r}^\pi \sum_{x \in U, x \in \omega(r)} P_x(1)P_x(0)\mathbf{k}_i^2(x)g_m^\pi(x)$$
$$+ \frac{\gamma}{\Pi}\sum_{\pi=1}^{\prod}\sum_{m=1}^M\sum_{r=1}^R v_{m,r}^\pi \sum_{x \in U, x \in \omega(r)} P_x^2(1)\mathbf{k}_i^2(x)g_m^\pi(x). \tag{19}$$

When $||\nabla_i C - h_i\boldsymbol{\alpha}_i||_2 \leq \lambda$, $\boldsymbol{\alpha}_i(t+1) = \mathbf{0}$. Otherwise

$$\mathbf{d}_i = -\mathbf{H}_i^{-1}\left[\nabla_i C - \lambda\frac{\nabla_i C - h_i\boldsymbol{\alpha}_i}{||\nabla_i C - h_i\boldsymbol{\alpha}_i||_2}\right]. \tag{20}$$

Then $\boldsymbol{\alpha}_i$ is updated by $\boldsymbol{\alpha}_i(t+1) = \boldsymbol{\alpha}_i(t) + a(t)\mathbf{d}_i$. $a(t)$ is the step size determined by Armijo rule which satisfies

$$a(t) = \max[\delta^0, \delta^1, \ldots, \delta^l], 0 < \delta < 1, l > 0$$
$$s.t. \quad Q_i(\boldsymbol{\alpha}_i + a(t) \cdot \mathbf{d}_i) - Q_i(\boldsymbol{\alpha}_i) \leq \sigma a(t)||\nabla Q_i|| \tag{21}$$

where $||\nabla Q_i|| = -\nabla C_i \cdot \mathbf{d}_i + \lambda(||\boldsymbol{\alpha}_i + \mathbf{d}_i||_2 - ||\boldsymbol{\alpha}_i||_2)$. We set $l = 20, \sigma = 0.618$, and $\delta = 0.5$, ensuring both efficiency and search precision. The un-regularized bias is directly optimized by

$$d_0 = -\frac{\nabla_0 C}{\nabla_{00} C}, \ \alpha_0(t+1) = \alpha_0(t) + d_0. \tag{22}$$

In the original BCGD approach [34], each coefficient group is optimized cyclically with a determined order. However, we have found two techniques that accelerate BCGD significantly.

Firstly, in many real-world situations, the ratio of positive data is usually much smaller than that of negative data. According to our observation, the ratio of non-zero $\boldsymbol{\alpha}_i$ for positive data will be much larger than the ratio of negative data. Therefore, more computation on kernel coefficients corresponding to the positive labeled data not only makes the model more robust, but also makes the convergence procedure much faster. To take advantage of this knowledge, we heuristically re-allocate the computational resources by updating the coefficients of negative data every Q rounds and update the coefficients of positive data each round. In this paper, we find the following reasonable setting [40]:

$$Q = \left\lfloor \frac{n_-}{4n_+} \right\rfloor \tag{23}$$

where n_+ and n_- denotes the number of positive and negative labeled data. Secondly, for coordinate descent based method, previous study in SVM [13] experimentally showed that random permutation of sub-problems is effective to improve the convergence rate. We find that it can also accelerate S³MKL optimization procedure. By combining the two techniques, our new BCGD method is much faster than the original approach [34]. We outline the procedure in Algorithm 1, and the properties of our method will be discussed in the experiment.

Algorithm 1: Optimization Procedure for S³MKL

1: **Initialize** $\alpha_0(0)$, $\boldsymbol{\alpha}(0)$, $t = 1$.

2: **while** Stop Criterion not meet and $t <= t_{\max}$ **do**

3:     Randomly permute $(1, \ldots |L|) \to (\varphi(1), \ldots, \varphi(|L|)$

4:     **For** each $\boldsymbol{\alpha}_i$, $i \in (\varphi(1), \ldots, \varphi(|L|)$

5:         **If** $y(i) = 1$ or $(y(i) = -1$ && $\mod(t, Q) = 0)$

6:             Compute $\mathbf{H}_i$, C and $\nabla_i C$

7:             **If** $||\nabla_i C - h_i \boldsymbol{\alpha}_i(t)||_2 < \lambda$

8:                 Set $\boldsymbol{\alpha}_i(t+1) = 0$

9:             **else**

10:                 Get optimal $\mathbf{d}_i(t)$, and a^t using Armijo line search

11:                 Set $\boldsymbol{\alpha}_i(t+1) = \boldsymbol{\alpha}_i(t) + a(t)\mathbf{d}_i$

12:             **end if**

13:         **end For**

14:         Update $\boldsymbol{\alpha}_0(t)$ by **(22)**, $t = t + 1$.

15: **end while**

16: **Output:** classification model $\mathbf{f}$ with coefficients $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}$

G. Memory Consumption

We provide some comparison on the memory consumption of our method with [35] from two aspects. Firstly, our method only requires memory size of $O(M \cdot |L| \cdot |U|)$ to load all the kernel matrix, where for [35] the total size of $O(MN^2)$ is required. Secondly, the coordinate gradient descent method requires very little memory consumption for each iteration since it does not need to conduct any matrix inverse operation. As for [35], the expensive $O(N^3)$ matrix inverse operation is required for optimization and another $O(N^2)$ space is requires to store the inverse matrix. The computation complexity would be lower if the kernel matrices are sparse. However, the kernel sparsity does not always hold in visual learning domain.

IV. HASHING

A. MKLSH

Nearest neighbor search is one of the key components in modern information retrieval system. Among the relevant researches, locality sensitive hashing (LSH) [6] is a well-known method which performs probabilistic dimension reduction and approximated nearest neighbor search for high-dimensional data. The basic idea is to hash the input items so that similar items are mapped to the same bucket with high probability. Based on LSH, the nearest neighbors of the query could be approximately identified very quickly. Among many versions of LSH, we adopt a recent developed kernel version of LSH [16]. The intuitive of KLSH is performing LSH in high dimensional

Fig. 2. Hashing system built by using MKLSH. Images in the red dashed box and purple dashed box denote the ImageNet data and Web data, respectively. The bottom image is the query input which can be the labeled training image. D_H is the Hamming distance of the binary codes. The images marked by "red $\sqrt{}$" is the returned images.

kernel Hilbert space with respect to the specific kernel. The kernel LSH is written as

$$h(\phi(x)) = sign\left(\sum_{i=1}^{P} \mathbf{w}(i) K(x_i, x)\right) \tag{24}$$

where $\phi(x)$ denotes the implicit feature representation in the kernel Hilbert space of K. The weight vector $\mathbf{w}$ is calculated by some random sample selection and operation similar with kernel PCA. The readers can refer to [16] for more details.

Since the original KLSH is built based on some specific kernel, the retrieved samples are highly dependent on the discrimination power of this kernel. To enhance the system capacity under the scenario of real-world image processing, we propose to build a hashing system using multiple kernels. It is analogous to a nearest neighbor classifier voting, which could reduce the search variance by classifier ensemble. Suppose we build the hashing system by using M kernels. For each kernel we generate H hash functions, As a result, a set of MH dimensional binary codes are generated for each image in the database and the query images. Distance calculation is directly done on calculating the Hamming distance of the MH length codes. When searching for a query's nearest neighbors, the system only checks those images with the binary codes that have the top B minimum hamming distances to the code of the query image. To ensure that the approximated neighbor search has better recall rate, we repeatedly generate 3 hash tables, and then all the examples retrieved by each table are put together for each query as the resulting nearest neighbor sets.

The system is presented in Fig. 2, where we use different color to represent KLSH using different kernels. The kernels used for constructing the hashing system are listed in Table V. Four kernels are used in this experiment. We set $H = 64$, which means a 64-bit hash function for each kernel, and 256-bit hash function for MKLSH. The images whose binary codes are most similar to the query are returned by the system. We will explain how to use this system for two different applications in Sections V–VII.

Data sources
- PASCAL VOC 2007 [9]: Visual Object Class Challenge. Training-Validation/Testing: 5,011/4,952.
- ImageNet [7]: subset of 20 classes of the large scale ground truth image data corpus. Total number: 21,500.
- Web data: Web image dataset downloaded from the Web with 250 queries. Total number: 2,130,428.

Environment

OS: Windows XP; Computer: Dell Optiplex 745 desktop; CPU: Intel(R) Core Duo E6300 @1.86GHz; Memory: 3.0G RAM; Programming language: Microsoft Visual Studio 2008

(a)

(b)

Fig. 3. (a) Framework of image classification. (b) Framework of personalized image re-ranking.

B. Clustering Based on Hashing Output

As has been discussed in Section III-D, to divide the training data into several clusters, the most common way is to conduct K-means based on the original feature representation. However, this not only requires extra memory to store the original feature, but also leads to extra time for clustering. Specifically, in our method, we hash all the labeled images and unlabeled images in the database, so each image will have a corresponding binary hash code with length MH. We directly conduct integer K-means on the binary codes.

Compared with clustering using the original feature, clustering using the hash code only requires very small extra cost because it avoids data exchange with the hard disk since the hash codes are always kept in the memory.

V. APPLICATIONS

We apply our method on two image applications, automatic image classification and personalized image re-ranking based on the hashing system in Fig. 2. To evaluate our method, we firstly construct an image database which includes three different image datasets as described in Table I. We use a challenging PASCAL VOC07 data as the labeled training data and test data for image classification. The ImageNet image subset is used for personalized image re-ranking. The large scale Web image data is used in both applications. We construct MKLSH on this database as in Fig. 2.

For image classification as in Fig. 3(a), the unlabeled data used for S³MKL training are Web images obtained by querying with all the labeled data. Both the labeled data and the chosen "informative" Web data are combined for S³MKL training.

As the second application, we consider personalized re-ranking as the technique that improves the results of the Web search to better fit the expectation of individual user. Therefore, some interaction should be involved, as we assume that user preference information has been encoded in the user labels. This is a little different from many previous re-ranking approaches where they only provide globally re-ranked results without user interaction. To better facilitate this need, we design and simulate the following active labeling process in Fig. 3(b).

Firstly, users input a query to the search engine. A set of Web images are returned. The interface provides users with some samples chosen randomly from the top N ranked images. Users label the images they prefer as relevant (1), and images they
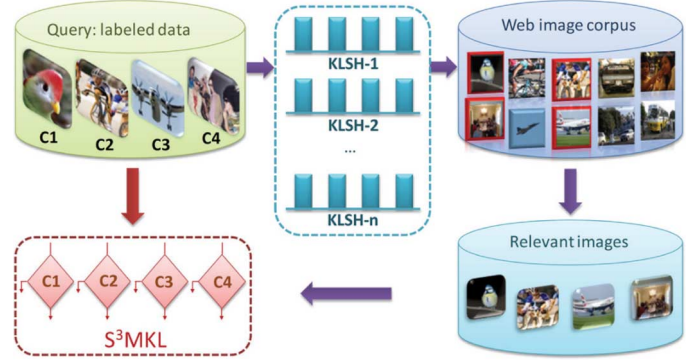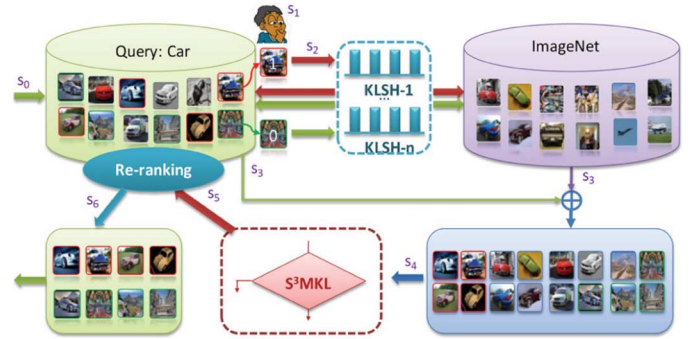
do not as irrelevant (0). Next, the user-labeled images are fit into the MKLSH system. As Fig. 2 shows, some images in ImageNet (in red dashed box) which are visually similar and semantically identical with the user-marked relevant images are returned as positive labeled images, and images from the other classes which are visually similar with the user marked irrelevant images are returned as negative labeled images. Additionally, some unlabeled Web images (in purple dashed box in Fig. 2) that are visually similar with the user labeled Web images are picked up as "pseudo" labeled data. Then the user labeled images, the chosen images from ImageNet, and the "pseudo" labeled images from the Web corpus are put together as the labeled training set, and all the other Web images are unlabeled training set. We conduct a transductive learning procedure by S³MKL on these image data. Finally all Web images are re-ranked according to the output of S³MKL.

The reason we choose images in ImageNet is that we need to augment the labeled data. Since users would not provide many labeled images, to propagate user's preference in a reasonable manner, nearest neighbor search is needed to prevent the user preference from being washed out by incorporating too many images. Secondly, the reason we choose "pseudo" labeled images from the returned Web images is that we believe that users may prefer some images they have not labeled in the Web images. Propagating the preference to some possible samples in the Web corpus would help to fully express the user intentions.

VI. EXPERIMENTS

We conduct experiments to testify our method. In Section VI-A, we conduct experiment on a machine learning dataset, the USPS data. In Section VI-B, we provide some analysis on the time complexity and scalability. In Section VI-C, we analyze the robustness of MKLSH. In Sections VI-D and VI-E, we present experiments on image annotation. Finally, in Section VI-F, we conduct experiment on personalized image re-ranking based on S^3MKL and MKLSH. The experimental settings are shown in Table I. All the features are pre-computed and the evaluation on training time does not include the time of feature computation.

From Section III we see that there are mainly three important parameters, γ, λ, and R. To conduct a parameter tuning process, we split the USPS training data into 5 subsets, and 5-folds cross validation are run on these subsets. For VOC'07 dataset, we employ the parameter tuning on the training/validation split provided by the VOC'07 official organizers [9]. The parameter setting of the three parameters described in Section III is determined by such a validation process so as to ensure that our model achieves both effectiveness and efficiency. Finally, we use both training and validation data for model training, and the experimental results reported for the rest of this paper are evaluated on the test data of all the datasets.

A. Handwritten Digit Recognition

The USPS dataset contains 10 handwritten digits, with 7291 training data and 2007 testing data. Excellent performance was achieved on this dataset using classifier combination and specific distance measure such as the tangent distance. We empirically generate 5 ordinary kernels using inner product, polynomial kernel and RBF kernel with 3 different bandwidths. We randomly choose part of the training data as the labeled samples, and the rest as the unlabeled data. The methods we evaluate in this part of experiment are:

- $S^3MKL(M)$: our approach (Section III-C and [40]) using M kernels ($M \leq 5$).
- $GS^3MKL(M)$: our group-wise approach (Section III-D) using M kernels ($M \leq 5$) and $R = 10$. K-means is used to divide the training data into clusters on the original feature.
- HFM: method in [5] using 5 kernels and $\lambda = 0.02$.
- SSLMN: SSL on multiple networks [35] using 5 kernels and $c = 10$.
- S^3VM: Semi-supervised SVM [3] using average kernel from 5 kernels and $C = 100$.

Since our S^3MKL and GS^3MKL are exactly the same as HFM when $\gamma = 0$, we use the optimization method proposed in this paper to minimizing the objective function of HFM, where the only difference is the calculation of gradient and Hessian for each sub-problem. For SSLMN, we use a classical gradient descent method on the dual problem as described in [35], finding the optimal kernel combination coefficients. The final prediction for all the unlabeled data is calculated by (10) in [35]. For S^3VM, we solve the objective function with SVM-light package [15].

For SSLMN and S^3VM, we merge the unlabeled training data and the test data into one set as the unlabeled data and we only
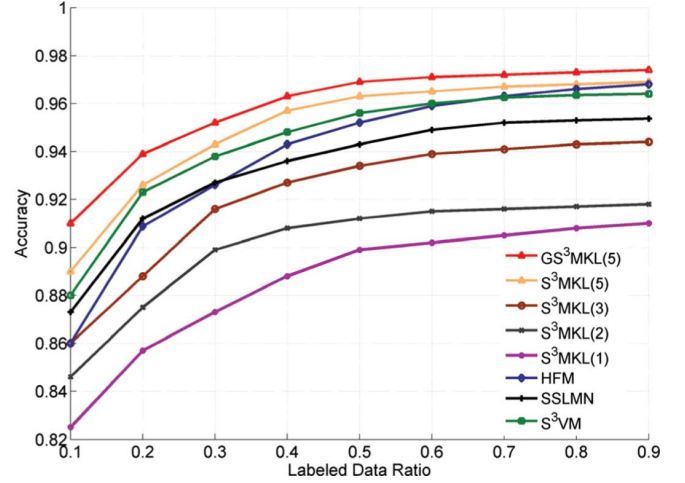

Fig. 4. Classification accuracy with different labeled data ratio on USPS dataset.

evaluate the classification accuracy on the test data so as to make the result comparable with other inductive approaches. For our method, the setting of parameter γ and λ is not very sensitive, we set $\gamma = 0.2 \cdot |L|$ and $\lambda = 0.02$. We repeat the experiments 10 times, and the mean accuracy under different labeled data ratio is shown in Fig. 4.

From the performance curves of S^3MKL using different number of kernels, we find that our method is capable of taking advantage of multiple kernels, as the accuracy increases when more kernels are incorporated. Furthermore, by incorporating group-wise *CEC*, our method GS^3MKL achieves a higher performance because it better fits the reference condition expectation on each data cluster.

When the ratio of the labeled data is small, our methods $S^3MKL(5)$ and $GS^3MKL(5)$ significantly outperform other methods. When the ratio of labeled data is large, our methods achieve remarkable prediction power compared with the supervised logistic regression [5] due to the regularization on the conditional expectation on the unlabeled data. When the size of the labeled data increases, the performance of our approach and [5] will be likely to converge, as the influence of the regularization of the unlabeled data reduces.

We also observe that our methods outperform [35], since the sample specific kernel weight assignment provides more deliberate fusion capability than the uniform kernel weight assignment. Our methods also outperform S^3VM using average kernel for similar reasons.

B. Time Complexity and Scalability

We evaluate the time complexity of our approach on USPS dataset. To study the convergence property of our optimization method based on BCGD, we sample 10% and the other 90% of the training samples as the labeled data and the unlabeled data for our approach. As has been introduced in the previous section, we improve the original BCGD approach by incorporating two techniques, namely, heuristic computation reallocation using the data imbalance and random permutation of sub-problems. In Fig. 5, we denote the original BCGD, BCGD
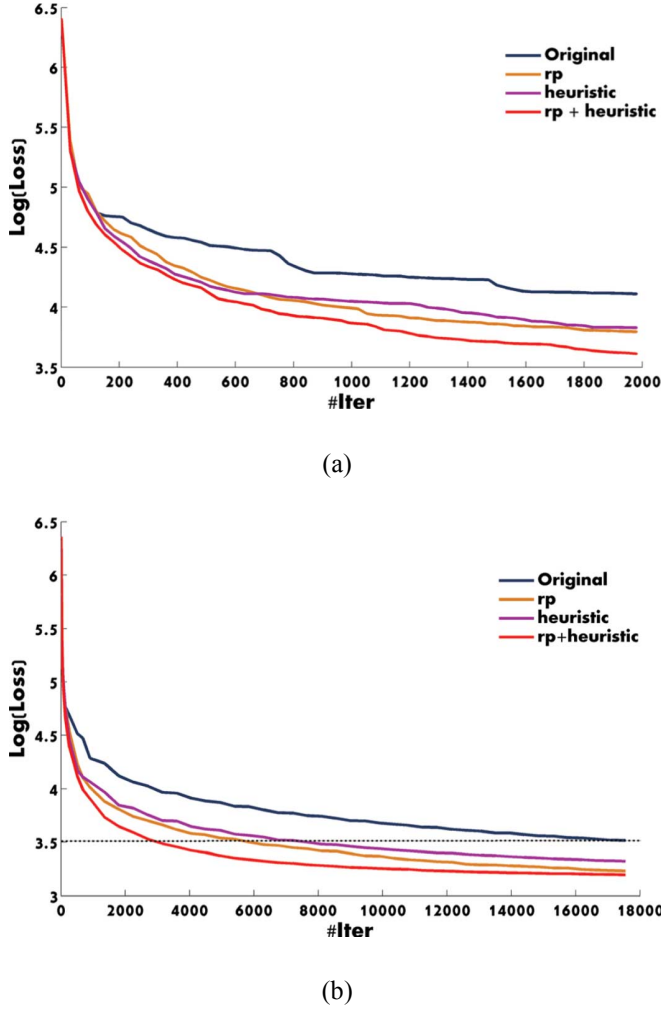
(a)

(b)

Fig. 5. (a) Value of objective function in the first 2000 rounds of scanning coefficient groups. (b) Whole convergence curves.

using computation re-allocation [40], BCGD using random permutation, and BCGD using both as *original*, *heuristic*, *rp*, and *rp+heuristic*, respectively. To fairly compute the four methods, we record the values of objective function with respect to the cumulative number of scanning of each coefficient group α_i. The curves in Fig. 5 show how these two techniques help to accelerate the training procedure. We see from Fig. 5(a) that all four methods descend fast at the beginning of the iteration, but they show different properties. Firstly, for the original BCGD, we observe that the convergence curve declines like a ladder shape, which means it has to take several steps to find a coefficient group α_i that can lead to the decline of objective function. By using either technique of the two, the curves descend more quickly and smoothly because the average time to find a working coefficient group is reduced.

Secondly, from Fig. 5(b), we further see that random permutation is a very effective way to ensure the model is more sufficiently optimized, which can be observed from the objective values after 10 K iteration. Incorporating computation reallocation leads to fast decline at the beginning. On the whole, incorporating both techniques leads to about 6 times acceleration compared with the original BCGD.

TABLE II
OVERALL TRAINING TIME STATISTICS

Method	S³MKL	GS³MKL	SSLMN	S³VM
Time (s)	183	201	2503	2794

TABLE III
SCALABILITY EVALUATION WITH DIFFERENT TRAINING SIZE

Ratio	0.2	0.5	0.8	1
S³MKL	65/0.61	108/0.42	152/0.34	183/0.30
GS³MKL	75/0.64	122/0.44	172/0.35	201/0.32

To compare the training efficiency of our method with other approaches, we calculate the overall training time in Table II. Our methods S^3MKL and GS^3MKL ($R = 10$) are at least 10 times faster than other approach. GS^3MKL spends a little more time than S^3MKL because it will take more time to calculate the model expectation on each data cluster. SSLMN takes a lot more time because matrix inversion has to be taken for each step of the gradient calculation. In addition, for S^3VM, an integer programming is necessary to guess the labels of the unlabeled data. Both of the optimization method used in SSLMN and S^3VM will be prohibitive for large scale data processing.

Moreover, we conduct some scalability analysis on how our approaches adapt with the training size growth. We randomly sample some subsets from the USPS training data, keeping the labeled versus unlabeled ratio as 1:9 consistently. The training time (the left of the backslash) and the ratio of non-zero elements (the right of the backslash) are recorded in Table III. From the statistics, we can see that when the training size grows, the ratio of the non-zero support vectors tends to be reduced. The reasons may come from two sides. Firstly, the estimation of the reference expectation and model expectation tends to be more stable and more accurate. Therefore, the model can be less sensitive to the outliers. Secondly, the number of non-zero support vectors does not grow linearly with the training size, which validates that our methods can effectively controls the model complexity. As the ratio of non-zero elements reduces when we increase the training size, the optimization process can be further accelerated because more zero element can be quickly filtered out. Therefore, the training time grows sub-linearly with the growth of training size. From the above comparison, we declare that our method is more scalable.

C. Retrieving Images by MKLSH

To demonstrate the advantage of MKLSH, we conduct experiments on VOC 2007 dataset [9]. We use the trainval images of VOC as the database, and 1000 images are randomly chosen from the test set as queries. The returned items are those images in the top 3 nearest buckets and they are ranked by their average similarity calculated on the four kernels (Table V) with the queries. We compare MKLSH (4 kernels) with 2 kernel version (Gist and PHOG), one kernel version (PHOG), as well as the original LSH [6] using multiple features and single feature for comparison. For multiple feature version of LSH, we concatenate the features into one vector for simplicity, which is a common way for processing multiple features. Since there are

Fig. 6. Some results retrieved by MKLSH (4 kernels) on VOC 2007. The left images are the queries and the right images are returned results. The dots in red represent the images of the same class with the queries, and dots in green represent the images from different classes.

TABLE IV
AVERAGE PRECISION FOR MKLSH AND LSH

Method	Average Precision
MKLSH(4 kernels)	0.52
MKLSH(Gist and PHOG)	0.40
MKLSH(PHOG)	0.31
LSH (4 features)	0.42
LSH(PHOG)	0.27

many images containing multiple labels, the precision is measured as: $precision = n_c/(n_c + n_w)$, where n_c denotes the number of images with at least one class label identical with one of the class labels of the query, and n_w denotes the number of images whose class labels are different from the query.

The average precision of all the 1000 queries is shown in Table IV and some examples are demonstrated in Fig. 6. We can see that the average precision is improved when more kernels are incorporated. Compared with LSH, MKLSH also achieves a higher precision rate because the similarity calculated by using multiple kernels better models the inter-relation among images. Moreover, feature concatenating for LSH will also lead to extremely sparse data distribution and complicated inter-correlation among features because of different range of each feature entry.

For the first two queries in Fig. 6, we can see that, when the background is very clean or the object is large enough, the retrieved images are very relevant. Some query that returns bad result is also shown in Fig. 6, as shown in the third example. We notice that the returned images look very similar, but the visual context (background) of the images is much cluttered, which leads to the difficulty of identifying these samples. In this case, a hashing system with class specific kernel weight setting is a better choice. Another example which demonstrates how visual context improves the retrieve result is shown in the fourth row, where the query image is an indoor image that contains human, and the system returns a set of indoor images where many of them have one or more people. We will investigate how to take better advantage of the knowledge of visual context to enhance the robustness in future study.

D. Image Classification on VOC'07

We conduct image classification experiments on PASCAL VOC 2007 dataset according to the procedure in Fig. 3(a). All the images have been hashed with MKLSH as shown in Fig. 2.

TABLE V
DETAILS OF FEATURES AND KERNELS

Features and kernels used in SSL

➢ 3 level PHOG-180 with Gaussian + χ^2 distance.
➢ 4 × 4 Color moment with RBF kernels.
➢ 2 × 2 Local binary pattern with Gaussian + χ^2 distance.
➢ 3 level spatial pyramid feature on dense Self Similarity bag of visual word with Gaussian + χ^2 distance.
➢ Geometric blur.
➢ Gist descriptor with Gaussian + χ^2 distance.
➢ 3 level spatial pyramid feature on dense bag of visual words with Gaussian + χ^2 distance.
➢ 3 level spatial pyramid feature on dense bag of color visual words with Gaussian + χ^2 distance.

Kernel used in MKLSH

➢ 3 level PHOG-180 with Gaussian + χ^2 distance.
➢ 4 × 4 Color moment with RBF kernels.
➢ 2 × 2 Local binary pattern with Gaussian + χ^2 distance.
➢ Gist descriptor with Gaussian + χ^2 distance.

Therefore, for experiment on GS^3MKL and its variant, clustering method in Section IV-B can be used because each image has a 256-bit binary code.

Previously, various types of features/kernels have been studied and evaluated [5], [31], [36], [45], [46]. In general, a complete image feature set should be able to include as more characteristics of color, shape, and texture as possible. We conduct some preliminary experiments on these previously used features/kernels. We choose 8 kinds of features/kernels from them for learning and 4 kinds of features/kernels for MKLSH, which guarantees good performance as well as the time and storage efficiency of MKLSH. The details of these kernels are listed in Table V.

In this part, the labeled data and the unlabeled data are randomly chosen from the VOC'07 train-val data, and the test data is VOC'07 test data. The aim of this experiment is to show how our method performs image classification on traditional classification problems. We evaluate the following method:

- S^3MKL: our approach (Section III-C and [40]).
- GS^3MKL: our group-wise S^3MKL (Section III-D). $R = 10$.
- $S^3MKL + MKH$: our approach (Section III-C and [40]) using MKLSH for unlabeled data selection.
- $GS^3MKL + MKH$: our group-wise S^3MKL (Section III-D) using MKLSH for unlabeled data selection. $R = 10$.
- SSLMN: SSL on multiple networks [35] where $c = 20$.
- S^3VM: transductive SVM [3] using weighted combination. β_m are tuned by using VOC validation data, and $\beta_m \geq 0, \sum \beta_m = 1, C = 500$.

We run these methods for different labeled data sampling ratios on VOC 2007 trainval data. For S^3MKL and GS^3MKL, all the rest of VOC trainval set is used as the unlabeled data. For $S^3MKL + MKH$ and $GS^3MKL + MKH$, only the unlabeled training images selected by MKLSH are treated as the unlabeled data. For the transductive SSLMN and S^3VM, we merge the unlabeled data and test data as one unlabeled set, and only evaluate on the test data to make the results comparable with others. For all the experiments conducted on VOC 2007 in this section, the evaluation criterion is *mean average precision* (*MAP*) [9]. Each
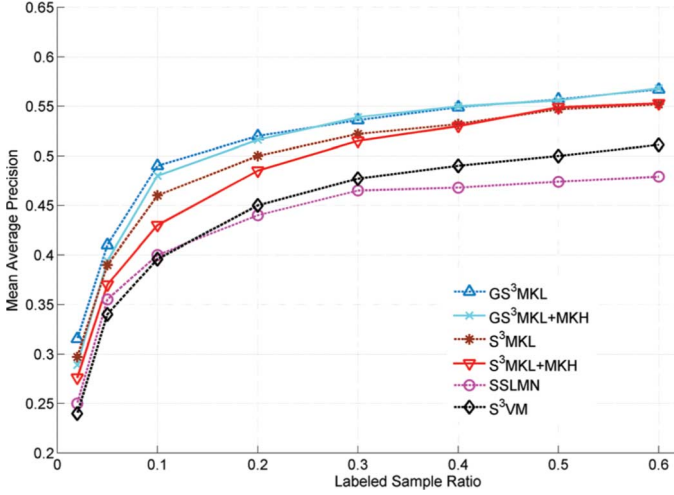
Fig. 7. Average precision with respect to the ratio of the labeled samples on VOC 2007 data.

TABLE VI
PERFORMANCE WITH DIFFERENT RATIO
OF SELECTED UNLABELED WEB IMAGES

Ratio	0.2	0.5	0.8	1
GS³MKL+MKH	0.566	0.579	0.584	0.586
S³MKL+MKH	0.559	0.565	0.569	0.570
S³VM+MKH	0.507	0.521	0.524	0.522
SSLMN+MKH	0.468	0.486	0.480	0.475

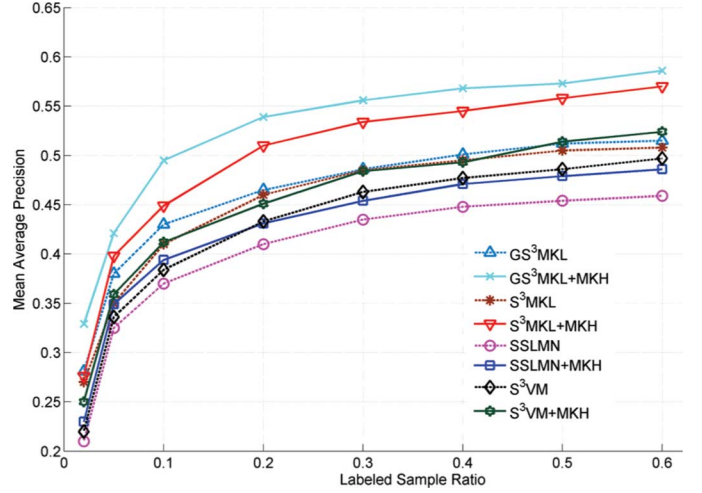Fig. 8. Average precision with respect to the ratio of the labeled samples on VOC 2007 data and Web data.

result is the average performance from the runs on 10 different random selections of labeled data. The performance with respect to different labeled data ratio is demonstrated in Fig. 7.

From this part of results, we see that our method S^3MKL and GS^{3}MKL outperforms SSLMN and S^{3}VM. GS^{3}MKL outperforms S^{3}MKL because the group-wise conditional expectation consensus better models the real-world image distribution. The effect of unlabeled sample selection is not very obvious for this part since the size of unlabeled data is small and the dataset is much cleaner compared with the real-world Web data. When the labeled data is small, using MKLSH will lead to the shortage of unlabeled data, which may explain why S^3MKL + MKH/GS^{3}MKL + MKH underperform S^{3}MKL/GS^{3}MKL when the labeled image ratio is small.

E. Image Classification Using Unlabeled Web Data

In this part, we conduct experiment on all methods as in Section VI-D, but the labeled data comes from a part of the trainval data of VOC 2007, and the unlabeled training data are the Web data. The evaluation set is VOC'07 test. The aim of this section is to show how our methods take advantage of the unconstrained unlabeled Web images. For S^3MKL, GS^{3}MKL, SSLMN, and S^{3}VM, we randomly choose an image subset from the Web image category whose class name is identical to the positive class as the unlabeled positive data, and randomly choose a subset from the rest of the Web images as the unlabeled negative data. We set the overall size of the randomly chosen unlabeled subset as 10 K. For S^3MKL + MKH and GS^{3}MKL + MKH, the unlabeled data are the subsets chosen by MKLSH as described in previous section. Since MKLSH is potentially applicable for all the other SSL methods, we also evaluate the two transductive baseline approaches with the unlabeled sample selected by MKLSH, and we denote them as S^3VM + MKH and SSLMN + MKH. As in Section VI-D, each result is the average performance from 10 runs.

Firstly, we figure out how MKLSH helps to avoid model degradation and enhance the model capacity under the unconstrained environment. We use 60% of the trainval data of VOC'07 as the labeled set and the queries of MKLSH. For

each query, MKLSH returns the images in the top 3 nearest buckets, where the average number of returned images is about 8. After removing duplicates, we have about 18 K Web images. We sample with a certain ratio from it as the unlabeled data for evaluating the four methods as recorded in Table VI. We see that our methods achieves the highest performance when we use all the unlabeled subset since increasing the number of unlabeled data can help to better estimate the model expectation. The other two baselines only achieve the highest performance when using a part of the subset. The reason for SSLMN + MKH may be that incorporating too many unlabeled data will sharply change the graph structure, while for S^3VM + MKH, there may be some disturbance. Therefore, for the rest of the evaluation, we use all subset for our approaches, while we only use 50% for SSLMN + MKH and 80% for S^3VM + MKH.

Secondly, by using the above mentioned sample selection schemes, the performance for all the methods with different labeled data ratio are demonstrated in Fig. 8. Our methods outperform other SSL approaches again. Specifically, GS^{3}MKL performs consistently better than other approaches without MKLSH, which prove the effectiveness of our method.

However, when we conduct semi-supervised learning using real-world image data, traditional SSL approaches are faced with great risk of model degradation without unlabeled data selection. This can be seen from the fact that the all methods without MKLSH in Fig. 8 performs inferiorly compared with Fig. 7. In contrast, by incorporating MKLSH, the performances of all the methods are significantly boosted. The influence of MKLSH can be analyzed from two aspects. Firstly, when the number of labeled samples is small, MKLSH identifies those images which are visually similar with the labeled images but

TABLE VII
AVERAGE PRECISION OF ALL THE METHODS WHEN THE LABELED TRAINING
DATA RATIO IS 0.6

	The source of unlabeled data	
	VOC'07	Web data
GS^{3}MKL+MKH	0.569	**0.586**
S^{3}MKL+MKH	0.553	**0.570**
GS^{3}MKL	**0.567**	0.515
S^{3}MKL	**0.552**	0.508
SSLMN+MKH	--	0.486
S^{3}VM+MKH	--	0.524
SSLMN	**0.479**	0.459
S^{3}VM	**0.511**	0.497

TABLE VIII
EVALUATION OF 20 QUERIES

Method	MAP
Initial rank	0.41
S^{3}MKL Re-ranking	0.74

contains more abundant semantic content. Secondly, when the sample ratio increases, the hashing system not only identifies those most "informative" unlabeled images, but also begins to filter out those noisy images which are ubiquitous on the Web. The unlabeled sample selection is particularly important for increasing the reliability of the positive unlabeled data.

We record the average precision of all the methods tested in Sections VI-D and VI-E when the labeled data ratio is 0.6 in Table VII. It further verifies our previous discussion that using the real-world unlabeled image data without control is very risky, and unlabeled data selection is an indispensable component for enhancing the robustness of SSL models when we conduct semi-supervised learning on large scale real-world image data.

F. Personalized Image Re-Ranking

In this section, we present evaluation of personalized re-ranking whose procedure is performed according to the introduction in Section V. The data we use in this section contain the Web data and ImageNet data subset. We select 20 object queries which are same as the object classes in VOC 2007. To evaluate the method, we invite several non-expert users to take part in the test. We ask them to randomly pick up some images to mark. In addition, after re-ranking using our method, we ask the same people to provide a personalized 4 level ground truth of the re-ranking results, namely, *favorite*, *acceptable*, *relevant* and *junk*. We treat the annotation of *favorite*, *acceptable*, and *relevant* as 1, and *junk* as 0. The performance measured in *MAP* is shown in Table VIII. After the re-ranking procedure using our method, we see that the score after re-ranking has been improved significantly compared with the initial rank returned by the Web in the sense of personal satisfaction.

Some re-ranking examples are shown in Fig. 9. The arrows stand for the samples that users choose to annotate. The red arrows point to the images that users mark as relevant (1), and the green ones as irrelevant (0). The dots on the images are the 4 level ground truth annotation provided by the users after the re-ranking procedure. We see from Fig. 9 that our method achieves significant ranking improvements in the sense of personalized user experience. For example, in the first results, the

Fig. 9. Personalized image re-ranking. The top rows are initial top results. The second rows are selected images from ImageNet. The last row is the top images of re-ranking. (Red: *favorite*; orange: *acceptable*; yellow: *relevant*; green: *junk*).

user picks an airplane landing images as his/her favorite images. The returned images contain more airplane landing images. Similar results can be observed for other data queries. We believe the result provides an alternative view for the study of Web image re-ranking.

VII. DISCUSSIONS AND CONCLUSIONS

We address two critical problems in real-world image data mining. Firstly, we propose a scalable semi-supervised multiple kernel learning approach, which makes use of multiple features/kernels to enhance the generalization power of semi-supervised learning. Secondly, to conduct training image selection in order to apply semi-supervised learning for Web image applications, we propose MKLSH, which combines multiple kernels for conducting effective approximated nearest neighbor search. Experimental results prove that our approach is promising for image annotation and personalized re-ranking.

Although promising results are obtained, our method has not solved all the problems. Firstly, from (19) we can see that the Hessian of C is not guaranteed to be positive definite. Since we use a second order approximation of C, the global convergence is guaranteed only when C is convex or quadratic [34]. Therefore, it will be a little risky when we use a positive diagonal Hessian when its actual value is negative. We observe in several

trials that the cost is stuck and vibrating around certain values, but this only lasts for several iterations. This situation is more likely to occur when γ is large, e.g., $\gamma = 10 \cdot |L|$. Fortunately, we do not need to impose heavy penalty on the unlabeled data, where a small value of γ as in Section III-C is enough for a good solution. Despite of this, in future work we will investigate how to make the optimization more robust.

Secondly, in MKLSH system we treat different kernels equally. However, we notice that different genres of images have different responses to the features. Therefore, the performance could be further improved by adaptive weight determination. It will be interesting to study hashing with multiple features in the future.

REFERENCES

[1] F. Bach, "Consistency of the group Lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, 2008.

[2] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. ICML*, 2004.

[3] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proc. NIPS*, 1998.

[4] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Computational Learning Theory*, 1998, pp. 92–100.

[5] L. Cao, J. Luo, F. Liang, and T. S. Huang, "Heterogeneous feature machine for visual recognition," in *Proc. ICCV*, 2009.

[6] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Computational Geometry*, 2004, pp. 253–262.

[7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[8] L. Duan, D. Xu, I. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *Proc. CVPR*, 2010.

[9] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool, The Pascal Visual Object Classes Challenge 2007 Results, 2007, Tech. rep.

[10] J. Fan, Y. Gao, and H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic concepts," in *Proc. ACM Multimedia*, 2004.

[11] M. Gonen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. ICML*, 2008.

[12] Y. Grandvalet and Y. Bengio, "Semi-supervised learning with entropy regularization," in *Proc. NIPS*, 2005.

[13] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large scale linear SVM," in *Proc. ICML*, 2008.

[14] J. Huang and T. Zhang, The Benefit of Group Sparsity, 2009, Tech. rep., arXiv: 0901.2962.

[15] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. ICML*, 1999.

[16] B. Kulis and K. Grauman, "Kernelized locality sensitive hashing for scalable image search," in *Proc. ICCV*, 2009.

[17] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2143–2157, Dec. 2009.

[18] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semi-definite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[19] G. Mann and A. McCallum, "Simple, Robust, Scalable Semi-Supervised learning via expectation regularization," in *Proc. ICML*, 2007.

[20] A. McCallum, G. Mann, and G. Druck, Generalized Expectation Criteria, Univ. Massachusetts, Amherst, 2007, Tech. Rep. 2007-60.

[21] L. Meier, S. van de Geer, and P. Buhlmann, "The group Lasso for logistic regression," *J. Royal Statist. Soc.: Series B*, vol. 70, no. 1, pp. 53–71, 2008.

[22] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. J. Zhang, "Correlative multi-label video annotation," in *Proc. ACM Multimedia*, 2007.

[23] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simple MKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008.

[24] A. Saffari, C. Leistner, and H. Bischof, "Regularized multi-class Semi-Supervised boosting," in *Proc. CVPR*, 2009.

[25] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: From transductive to semi-supervised learning," in *Proc. ICML*, 2005.

[26] V. Sindhwani and D. S. Rosenberg, "An RKHS for multi-view learning and manifold co-regularization," in *Proc. ICML*, 2008.

[27] A. Singh, R. D. Nowak, and X. Zhu, "Unlabeled data: Now it helps, now it doesn't," in *Proc. NIPS*, 2008.

[28] C. G. M. Snoek and M. Worring, "Early versus late fusion in semantic video analysis," in *Proc. ACM Multimedia*, 2005.

[29] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006.

[30] J. Tang, X. Hua, G. Qi, M. Wang, T. Mei, and X. Wu, "Structure sensitive manifold ranking for video concept detection," in *Proc. ACM Multimedia*, 2007.

[31] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[32] A. Torralba, K. P. Murphy, and W. T. Freeeman, "Sharing features: Efficient boosting procedures for multiclass object detection," in *Proc. CVPR*, 2004.

[33] A. Torralba, B. C. Russell, and J. Yuen, LabelMe: Online Image Annotation and Applications, MIT CSAIL, 2009, Tech. Rep.

[34] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Math. Program. B*, vol. 117, pp. 1–2, 2009.

[35] K. Tsuda, H. J. Shin, and B. Scholkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. 2, pp. ii59–ii65, 2005.

[36] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. ICCV*, 2007.

[37] V. N. Vapnik, *The Nature of Statistical Learning*. New York: Springer, 1995.

[38] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "Multiple kernel learning with high order kernels," in *Proc. ICPR*, 2010.

[39] S. Wang, Q. Huang, S. Jiang, and Q. Tian, "Nearest-neighbor classification using unlabeled data for real world image application," in *Proc. ACM Multimedia*, 2010.

[40] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "S³MKL: Scalable Semi-Supervised multiple kernel learning for image data mining," in *Proc. ACM Multimedia*, 2010.

[41] X. J. Wang, L. Zhang, F. Jing, and W. Y. Ma, "AnnoSearch: Image auto-annotation by search," in *Proc. CVPR*, 2006.

[42] L. Wu, S. C. H. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging," in *Proc. ACM Multimedia*, 2009.

[43] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proc. ACM KDD*, 2007.

[44] R. Yan, A. Natsev, and M. Campbell, "A learning-Based hybrid tagging and browsing approach for efficient manual image annotation," in *Proc. CVPR*, 2008.

[45] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Group sensitive multiple kernel learning for object categorization," in *Proc. ICCV*, 2009.

[46] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. CVPR*, 2006.

[47] X. Zhu, Semi-Supervised Learning Literature Survey, Univ. Wisconsin–Madison, 2006, Tech. Rep. 1530.

[48] X. Zhu and Z. Ghahramani, "Semi-Supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, 2003.

[49] D. Rosenberg, V. Sindhwani, P. Bartlett, and P. Niyogi, "A kernel for semi-supervised learning with multi-view point cloud regularization," *IEEE Signal Process. Mag.*, 2009.

[50] H. Hino and N. Murata, "A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning," *Neural Computat.*, vol. 22, no. 11, pp. 2887–2923, 2010.

[51] H. Hino, N. Reyhani, and N. Murata, "Multiple kernel learning by conditional entropy minimization," in *Proc. ICMLA*, 2010, pp. 223–228.
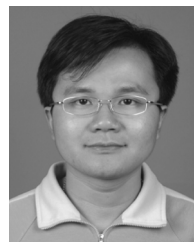
Shuhui Wang (S'12) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006. He is currently pursuing the Ph.D. degree in the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing.

His research interests include semantic image analysis, image and video retrieval, and large-scale web multimedia data mining.

Qingming Huang (SM'08) received the B.S. degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the Graduate University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, CAS. He has authored or coauthored nearly 200 academic papers in prestigious international journals and conferences. His research areas include multimedia video analysis, video adaptation, image processing, computer vision, and pattern recognition.

Dr. Huang is a reviewer for the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON COMMUNICATIONS. He has served as program chair, track chair, and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, PSIVT, etc.

Shuqiang Jiang (SM'08) received the M.S. degree from the College of Information Science and Engineering, Shandong University of Science and Technology, Shandong, China, in 2000, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 90 papers on the related research topics.

Qi Tian (M'02–SM'03) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992 and the Ph.D. degree in electrical and computer engineering from the University of Illinois, Urbana-Champaign, in 2002.

He is currently an Associate Professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA). His research interests include multimedia information retrieval and computer vision. He has published over 150 refereed journal and conference papers. His research projects were funded by NSF, ARO, DHS, SALSI, CIAS, and UTSA and he also received faculty research awards from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Labs. He took a one-year faculty leave at Microsoft Research Asia (MSRA) during 2008–2009.

Dr. Tian was the author of a Top 10% Best Paper Award in MMSP 2011, a Best Student Paper in ICASSP 2006, and a Best Paper Candidate in PCM 2007. He received the 2010 ACM Service Award. He has been serving as Program Chairs, Organization Committee Members, and TPCs for numerous IEEE and ACM Conferences, including ACM Multimedia, SIGIR, ICCV, ICME, etc. He is a Guest Editors of the IEEE TRANSACTIONS ON MULTIMEDIA, *Journal of Computer Vision and Image Understanding*, *Pattern Recognition Letters*, *EURASIP Journal on Advances in Signal Processing*, and *Journal of Visual Communication and Image Representation*, and is in the Editorial Board of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Journal of Multimedia* (JMM), and *Journal of Machine Visions and Applications* (MVA). He is a Member of ACM.