Multiview Vector-Valued Manifold Regularization for Multilabel Image Classification

Yong Luo, Dacheng Tao, *Senior Member, IEEE*, Chang Xu, Chao Xu, Hong Liu, and Yonggang Wen, *Member, IEEE*

***Abstract*— In computer vision, image datasets used for classification are naturally associated with multiple labels and comprised of multiple views, because each image may contain several objects (e.g., pedestrian, bicycle, and tree) and is properly characterized by multiple visual features (e.g., color, texture, and shape). Currently, available tools ignore either the label relationship or the view complementarily. Motivated by the success of the vector-valued function that constructs matrix-valued kernels to explore the multilabel structure in the output space, we introduce multiview vector-valued manifold regularization (MV^{3}MR) to integrate multiple features. MV^{3}MR exploits the complementary property of different features and discovers the intrinsic local geometry of the compact support shared by different features under the theme of manifold regularization. We conduct extensive experiments on two challenging, but popular, datasets, PASCAL VOC' 07 and MIR Flickr, and validate the effectiveness of the proposed MV^{3}MR for image classification.**

***Index Terms*— Image classification, manifold, multilabel, multiview, semisupervised.**

I. Introduction

A NATURAL image can be summarized by several keywords or labels. To conduct image classification by directly using binary classification methods [1], [2], it is necessary to assume that labels are independent, although most labels appearing in one image are related to one another. Examples are given in Fig. 1, where A1–A3 shows a person riding a motorbike, B1–B3 indicates sea usually co-occurring with sky, and C1–C3 shows some clouds in the sky. This multilabel nature makes image classification intrinsically different from simple binary classification.

Manuscript received June 16, 2012; revised December 31, 2012; accepted December 31, 2012. Date of publication February 8, 2013; date of current version March 8, 2013. This work was supported in part by NBRPC 2011CB302400, NSFC 60975014, 61121002, JCYJ20120614152136201, and NSFB 4102024, and in part by the Australian Research Council under Discovery Project DP-120103730.

Y. Luo, C. Xu, and C. Xu are with the Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: yluo180@gmail.com; changxu1989@gmail.com; xuchao@cis.pku.edu.cn).

D. Tao is with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering & Information Technology, University of Technology, Sydney, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

H. Liu is with the Engineering Laboratory on Intelligent Perception for Internet of Things, Shenzhen Graduate School, Peking University, Shenzhen 518055, China (e-mail: liuh@pkusz.edu.cn).

Y. Wen is with the Division of Networks and Distributed Systems, School of Computer Engineering, Nanyang Technological University, 639798, Singapore (e-mail: ygwen@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2238682

Moreover, different labels cannot be properly characterized by a single feature representation. For example, the color information (e.g., color histogram), shape cue (encoded in scale-invariant feature transform (SIFT) [3]), and global structure (e.g., GIST [4]) can effectively represent natural substances (e.g., sky, cloud, and plant life), man-made objects (e.g., aeroplane, motorbike, and TV monitor), and scenes (e.g., seaside and indoor), respectively, but cannot simultaneously illustrate all these concepts in an effective way. Each visual feature encodes a particular property of the images and characterizes a particular concept (label), so we treat each feature representation as a particular view for characterizing images. Fig. 1(a)–(c) indicates that SIFT representation is effective in describing a motorbike and GIST can capture the global structure of a person on the motorbike. Fig. 1(d)–(f) shows that GIST performs well in recognizing seaside scenes, while the color information can be used as a complementary aid for recognizing the blue seawater. From Fig. 1(g)–(i), we can see that RGB usually represents cloud well and GIST is helpful when RGB fails. For example, the RGB representations of C1 and C3 are not very similar but their GIST distance (0.22) is very small due to the sky scene structure. This multiview nature distinguishes image classification from single-view tasks, such as texture segmentation [5] and face recognition [6].

The vector-valued function [7] has recently been introduced to resolve multilabel classification [8] and has been demonstrated to be effective in semantic scene annotation. This method naturally incorporates the label dependencies into the classification model by first computing the graph Laplacian [9] of the output similarity graph, and then uses this graph to construct a vector-valued kernel. This model is superior to most of the existing multilabel learning methods [10]–[12] because it naturally considers the label correlations and efficiently outputs all the predicted labels at one time.

Although the vector-valued function is effective for general multilabel classification tasks, it cannot directly handle image classification problems that include images represented by multiview features. A popular solution is to concatenate all the features into a long vector. This concatenation strategy not only ignores the physical interpretations of different features but also encounters the overfitting problem given the limited training samples.

We thus introduce multikernel learning (MKL) to the vector-valued function and present a multiview vector-valued manifold regularization (MV^{3}MR) framework for handling the multiview features in multilabel image classification. MV^{3}MR associates each view with a particular kernel, assigns a higher

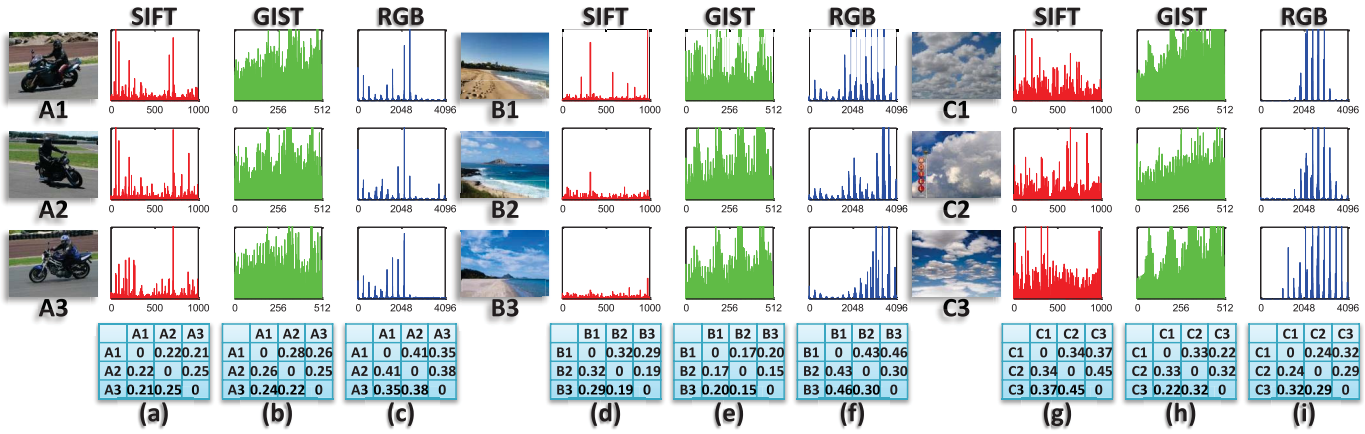


Fig. 1. A1–A3, B1–B3, and C1–C3 are images of a person riding a motorbike, a seaside, and clouds in the sky, respectively. (a)–(i) Feature representations and a distance matrix of the samples in a particular view. All the distances have been normalized here.

weight to the view/kernel carrying more discriminative information, and explores the complementary nature of different views.

In particular, MV^{3}MR assembles the multiview information through a large number of unlabeled images to discover the intrinsic geometry embedded in the high-dimensional ambient space of the compact support of the marginal distribution. The local geometry, approximated by the adjacency graphs induced from multiple kernels of all the corresponding views, is more reliable than that approximated by the adjacency graph induced from a particular kernel of any corresponding view. In this way, MV^{3}MR essentially improves the vector-valued function for multilabel image classification.

Because the hinge loss is more suitable for classification than the least squares loss [13]–[15], we derive an support vector machine (SVM) formulation of MV^{3}MR which results in a multiview vector-valued Laplacian SVM (MV^{3}LSVM). We carefully design the MV^{3}LSVM algorithm so that it determines the set of kernel weights in the learning process of the vector-valued function.

We thoroughly evaluate the proposed MV^{3}LSVM algorithm on two challenging datasets, PASCAL VOC' 07 (VOC) [16] and MIR Flickr (MIR) [17], by comparing it with a popular MKL algorithm [18], a recently proposed MKL method [19], and competitive multilabel learning algorithms for image classification, such as multilabel compressed sensing (MLCS) [20], canonical correlation analysis (CCA) [12], and vector-valued manifold regularization [8] in terms of mean average precision (mAP), mean area under curve (mAUC), and Ranking loss (RL). The experimental results suggest the effectiveness of MV^{3}LSVM.

The rest of this paper is organized as follows. Section II summarizes the recent work in multilabel learning, MKL and image classification. In Section III, we introduce manifold regularization and its vector-valued generalization. We depict the proposed MV^{3}MR framework and its SVM formulation in Section IV. Extensive experiments are presented in Section V and we conclude this paper in Section VI.

II. Related Work

A. Multilabel Learning

Multilabel classification has received intensive attention in recent years [21]–[23]. Some methods extend traditional multiclass algorithms to cope with the multilabel problem. AdaBoost.MH [24] adds the label value to the feature vector and then applies AdaBoost on weak classifiers. A ranking algorithm is presented in [25] by adopting the RL as the cost function in SVM. Multilabel k nearest neighbours (ML-KNN) [26] is an extension of the KNN algorithm to deal with multilabel data, and CCA has also recently been extended to the multilabel case by formulating it as a least-squares problem [12].

Other works concentrate on preprocessing the data so that standard binary or multiclass techniques can be utilized. For example, multiple labels of a sample belong to a subset of the whole label set and we can view this subset as a new class [27]. This may lead to a large number of classes and a more common strategy is to learn a binary classifier for each label [1], [2]. Considering that the labels are often sparse, a compressed sensing method is proposed for multilabel prediction [20].

Various approaches have been proposed to improve prediction accuracy by exploiting label correlations [8], [11], [28], [29]. Sun *et al.* [28] proposed the construction of a hypergraph to exploit the label dependencies. In [29], a common subspace is assumed to be shared among all labels, and the correlation information contained in different labels can be captured by learning this low-dimensional subspace. A max-margin method is proposed in [11], where prior knowledge of label correlations is incorporated explicitly in the multilabel classification model.

None of the approaches mentioned above considers the features to be used; however, an image with multiple labels usually indicates that it contains multiple objects. As far as we know, there is no single kind of feature that can describe a variety of objects very well. Therefore, how to combine different features is a critical issue in multilabel image classification and we consider MKL for this purpose in this paper.

B. Multikernel Learning

Classical kernel methods are usually based on a single kernel [30], [31]. MKL [32], in which a kernel-based classifier and a convex combination of the kernels are learned simultaneously, has attracted much attention. Lanckriet *et al*. [32] have introduced MKL for binary classification and solved it with semidefinite programming techniques. The MKL algorithm was further developed by Sonnenburg *et al*. [33] in the presentation of a semi-infinite linear program. In [18], MKL is reformulated by using a weighted L2-norm regularization to replace the mixed-norm regularization and adding an L1-norm constraint on the kernel weights. All of these MKL formulations are based on SVM and are not naturally designed for multilabel classification. The proposed MV^{3}MR framework extends MKL to handle the multilabel problem and model label interdependencies.

C. Image Classification

Image classification has been widely used in many computer-vision-related applications such as image retrieval and web content browsing. In recent years, more than a dozen methods have been proposed and representative works can be grouped into three categories.

1) *Single-View Learning for Image Classification:* This category contains many recent image classification schemes, e.g., dictionary learning [34] and spatial pyramid matching [35]. For example, Labusch *et al.* [36] proposed to integrate sparse-coding and local-maximum operation to extract local features for handwritten digit recognition. In [37], a nonlinear coding scheme was introduced for local descriptors such as SIFT. Yang *et al.* [38] explored the local co-occurrences of visual words over the spatial pyramid.
2) *Multiview Learning for Image Classification:* Multiview learning is an active current research topic [39]–[41]. Schemes in this category utilize the features from different views (or multiview features) to boost image classification performance. In this paper, the concept of "views" used for learning refers to different features or attributes for depicting the objects to be classified. It should be noted that for some other applications in vision and graphics, "views" mean different spatial viewpoints [42]–[44]. A semisupervised boosting algorithm is proposed in [45], in which images measured by different views are used to construct a prior and formulate a regularization term. Guillaumin *et al*. [2] combined 15 visual representations [e.g., SIFT, GIST, and hue, saturation and value (HSV)] with the tag feature for semisupervised image classification. Combining the visual and textual information has been utilized for clustering [46] and web page classification [47].
3) *Multilabel Learning for Image Classification:* This category is motivated by the success of multilabel learning and has demonstrated promising image classification performance. For example, Bucak *et al*. [48] proposed a ranking-based algorithm to tackle the multilabel problem with incompletely labeled data by introducing a group lasso regularizer in optimization. Unlike traditional multilabel methods that always consider positive label correlations, a novel approach is presented in [49] to make use of the negative relationship of categories.

Although it has been widely acknowledged that both multiview representation and label interdependencies are important for multilabel image classification, most of the existing approaches do not take both of them into consideration. Most existing multiview approaches assume that different views (features) contribute equally to label prediction. In contrast to these approaches, the proposed MV^{3}MR naturally explores both the complementary property of multiview features and the correlations of different labels under the manifold regularization scheme.

III. Manifold Regularization and Vector-Valued Generalization

This section briefly introduces the manifold regularization framework [9] and its vector-valued generalization [8]. Given a set of l labeled examples $D_l = \{(x_i, y_i)_{i=1}^{l}\}$ and a relatively large set of u unlabeled examples $D_u = \{(x_i)_{i=l+1}^{N=l+u}\}$, we consider a nonparametric estimation of a vector-valued function $f : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}^n$ and n is the number of labels. This setting includes $\mathcal{Y} = \mathbb{R}$ as a special case for regression and classification.

A. Manifold Regularization

Manifold learning has been widely used for capturing the local geometry [50] and conducting low-dimensional embedding [51], [52]. In manifold regularization, the data manifold is characterized by a nearest neighbor graph $\mathcal{W}$, which explores the geometric structure of the compact support of the marginal distribution. The Laplacian $\mathcal{L}$ of $\mathcal{W}$ and the prediction $\mathbf{f} = [f(x_1), \ldots, f(x_N)]$ are then formulated as a smoothness constraint $\|f\|_I^2 = \mathbf{f}^T \mathcal{L} \mathbf{f}$, where $\mathcal{L} = \mathcal{D} - \mathcal{W}$ and the diagonal matrix $\mathcal{D}$ is given by $\mathcal{D}_{ii} = \sum_{j=1}^{N} \mathcal{W}_{ij}$. The manifold regularization framework minimizes the regularized loss

$$\underset{f \in \mathcal{H}_k}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_k^2 + \gamma_I \|f\|_I^2 \qquad (1)$$

where L is a predefined loss function, k is the standard scalar-valued kernel, i.e., $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, and $\mathcal{H}_k$ is the associated reproducing kernel Hilbert space (RKHS). Here, γ_A and γ_I are trade-off parameters to control the complexities of f in the ambient space and the compact support of the marginal distribution. The representer theorem [9] ensures the solution of (1) takes the form $f^*(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$, where $\alpha_i \in \mathbb{R}$ is the coefficient. Since a pair of close samples means that the corresponding conditional distributions are similar, the manifold regularization $\|f\|_I^2$ helps the function learning.

B. Vector-Valued Manifold Regularization

In the vector-valued RKHS, where a kernel function K is defined and the corresponding $\mathcal{Y}$-valued RKHS is denoted by

$\mathcal{H}_K$, the optimization problem of the vector-valued manifold regularization (VVMR) is given by

$$\underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l}\sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_K^2 + \gamma_I \langle \mathbf{f}, \mathcal{M}\mathbf{f}\rangle_{\mathcal{Y}^{u+l}} \quad (2)$$

where $\mathcal{Y}^{u+l}$ is the $u+l$ direct product of $\mathcal{Y}$, and the inner product takes the form

$$\langle (y_1, \ldots, y_{u+l}), (w_1, \ldots, w_{u+l})\rangle_{\mathcal{Y}^{u+l}} = \sum_{i=1}^{u+l} \langle y_i, w_i\rangle_{\mathcal{Y}}.$$

The function prediction $\mathbf{f} = [f(x_1), \ldots, f(x_{u+l})] \in \mathcal{Y}^{u+l}$. The matrix $\mathcal{M}$ is a symmetric positive operator that satisfies $\langle \mathbf{y}, \mathcal{M}\mathbf{y}\rangle \geq 0$ for all $\mathbf{y} \in \mathcal{Y}^{u+l}$ and is chosen to be $\mathcal{L}\otimes I_n$. Here, $\mathcal{L}$ is the graph Laplacian, I_n is the $n\times n$ identity matrix, and $\otimes$ denotes the Kronecker (tensor) matrix product. For $\mathcal{Y} = \mathbb{R}^n$, an entry $K(x_i, x_j)$ of the $n\times n$ vector-valued kernel matrix is defined by

$$K(x_i, x_j) = k(x_i, x_j)(\gamma_O L_{\text{out}}^{\dagger} + (1-\gamma_O) I_n) \quad (3)$$

where $k(\cdot,\cdot)$ is a scalar-valued kernel, and $\gamma_O \in [0, 1]$ is a parameter. Here, $\mathcal{L}_{\text{out}}^{\dagger}$ is the pseudo-inverse of the output labels' graph Laplacian. The graph can be estimated by viewing each label as a vertex and using the nearest neighbors method. The representation of the jth label is the jth column in the label matrix $Y \in \mathbb{R}^{N\times n}$, in which $Y_{ij} = 1$ if the jth label is manually assigned to the ith sample, and -1 otherwise. For the unlabeled samples, $Y_{ij} = 0$.

It has been proved in [8] that the solution of the minimization problem (2) takes the form $f^*(x) = \sum_{i=1}^{N} K(x, x_i)a_i$. By choosing the regularization least squares (RLS) loss $L(f, x_i, y_i) = (f(x_i) - y_i)^2$, we can estimate the column vector $\mathbf{a} = \{a_1, \ldots, a_{u+l}\} \in \mathbb{R}^{n(u+l)}$ with each $a_i \in \mathcal{Y}$ by solving a the Sylvester equation

$$-\frac{1}{l\gamma_A}(J_l^N G^k + l\gamma_I \mathcal{L} G^k) A Q - A + \frac{1}{l\gamma_A} Y = 0 \quad (4)$$

where $a = \text{vec}(A^T)$ and $Q = (\gamma_O \mathcal{L}_{\text{out}}^{\dagger} + (1-\gamma_O) I_n)$ and J_l^N is a diagonal matrix with the first l entries 1 and the others 0. Here, G^k is the Gram matrix of the scalar-valued kernel k over the labeled and unlabeled data. We refer the reader to [8] for a detailed description of the vector-valued Laplacian RLS.

IV. Multiview Vector-Valued Manifold Regularization

To handle multiview multilabel image classification, we generalize VVMR and present MV^{3}MR. In contrast to [2], which assumes that different views contribute equally to the classification, MV^{3}MR assumes that different views contribute to the classification differently and learns the combination coefficients to integrate different views.

Fig. 2 gives an illustrative example which suggests that different views contribute to the classification differently and that learning the combination coefficients to integrate different views benefits the classification. Given five images from two classes, namely three cars of different colors (silvery white, blue, and red) and two different sky images, the optimal Gram matrix G_{opt}^k is shown on the right side for separating these images into two classes. On the left, there are four Gram matrices, which are two single Gram matrices G_1^k, G_2^k obtained from two different views, and their mean G_{avg}^k, as well as their linear combination G_{comb}^k with the learned coefficients. The figure indicates that G_{comb}^k is closer to the optimal Gram matrix G_{opt}^k than G_{avg}^k.

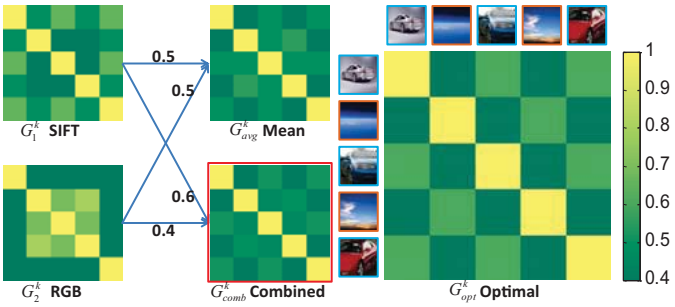


Fig. 2. Different views contributing to the classification differently. G_1^k is a Gram matrix constructed from SIFT [3]. G_2^k is obtained from RGB color histogram. G_1^k and G_2^k are complementary to each other. The learned linear combination G_{comb}^k of the two Gram matrices is closer to the optimal Gram matrix G_{opt}^k than the mean Gram matrix G_{avg}^k by simply averaging the two kernels.

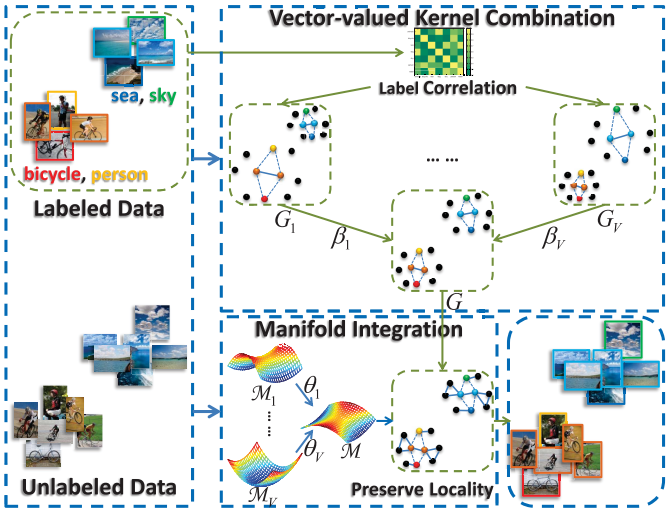


Fig. 3. Diagram of the proposed MV^{3}MR algorithm. The given labels are used to construct an output similarity graph, which encodes the label correlations. Features from different views of the labeled and unlabeled data are used to construct different Gram matrices (with label correlations incorporated) $G_v, v = 1, \ldots, V$ as well as the different graph Laplacians $\mathcal{M}_v, v = 1, \ldots, V$. We learn the weight β_v for G_v and θ_v for $\mathcal{M}_v$. The combined Gram matrix G is used for classification while preserving locality on the integrated manifold $\mathcal{M}$.

Given a small number of labeled samples and a relatively large number of unlabeled samples, MV^{3}MR first computes an output similarity graph by using the label information of the labeled samples. The Laplacian of the label graph is incorporated in the scalar-valued Gram matrix G_v^k over labeled and unlabeled data to enforce label correlations on each view, and the vector-valued Gram matrices $G_v = G_v^k \otimes Q$, $v = 1, \ldots, V$ can be obtained. Meanwhile, we also compute the vector-valued graph Laplacians $\mathcal{M}_v, v = 1, \ldots, V$ by using the features of the input data from different views. Then MV^{3}MR learns the kernel combination coefficient β_v for G_v

as well as the graph weight θ_v for $\mathcal{M}_v$ by the use of alternating optimization. Finally, the combined Gram matrix G together with the regularization on the combined manifold M is used for classification. Fig. 3 summarizes the above procedure. The technical details are given below.

A. Rationality

Let V be the number of views and v be the view index. On the feature space of each view, we define the corresponding positive-definite scalar-valued kernel k_v, which is associated with an RKHS $\mathcal{H}_{k_v}$. It follows from the functional framework [18] that, by introducing a nonnegative coefficient β_v, the Hilbert space $\mathcal{H}'_{k_v} = \{f | f \in \mathcal{H}_{k_v} : \frac{\|f\|_{\mathcal{H}_{k_v}}}{\beta_v} < \infty\}$ is an RKHS with kernel $k(x, x') = \beta_v k_v(x, x')$. If we define $\mathcal{H}_k$ as the direct sum of the space $\mathcal{H}'_{k_v}$, i.e., $\mathcal{H}_k = \oplus_{v=1}^{V} \mathcal{H}'_{k_v}$, then $\mathcal{H}_k$ is an RKHS associated with the kernel

$$k(x, x') = \sum_{v=1}^{V} \beta_v k_v(x, x'). \tag{5}$$

Thus, any function in $\mathcal{H}_k$ is a sum of functions belonging to $\mathcal{H}_{k_v}$. The vector-valued kernel $K(x, x') = k(x, x') \otimes Q = \sum_{v=1}^{V} \beta_v K_v(x, x')$, where we have used the bilinearity of the Kronecker product. Each $K_v(x, x') = k_v(x, x') \otimes Q$ corresponds to an RKHS according to the study of RKHS for the vector-valued functions [8]. Thus, the kernel K is associated with an RKHS $\mathcal{H}_K$. This functional framework motivates the MV^{3}MR framework. We will jointly learn the linear combination coefficients $\{\beta_v\}$ to integrate kernels for characterizing different views and the classifier coefficients $\{a_i\}$ in a single optimization problem. Moreover, to effectively utilize the unlabeled data, we construct graph Laplacians for different views and learn to combine all of them.

B. Problem Formulation

Under the multiview setting and the theme of manifold regularization, we propose to learn the vector-valued function f by linearly combining the kernels and graphs from different views. The optimization problem is given by

$$\begin{aligned} \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \quad & \frac{1}{l} \sum_{i=1}^{l} L(f, x_i, y_i) + \gamma_A \|f\|_K^2 + \gamma_I \langle \mathbf{f}, \mathcal{M}\mathbf{f} \rangle_{\mathcal{Y}^{u+l}} \\ & + \gamma_B \|\beta\|_2^2 + \gamma_C \|\theta\|_2^2 \\ \text{s.t.} \quad & \sum_{v=1}^{V} \beta_v = 1, \qquad \beta_v \geq 0 \\ & \sum_{v=1}^{V} \theta_v = 1, \qquad \theta_v \geq 0, v = 1, \ldots, V \end{aligned} \tag{6}$$

where $\beta = [\beta_1, \ldots, \beta_V]^T$ and $\theta = [\theta_1, \ldots, \theta_V]^T$. Both $\gamma_B > 0$ and $\gamma_C > 0$ are trade-off parameters. The decision function takes the form $f(x) + b = \sum_v f^v(x) + b$ and belongs to an RKHS $\mathcal{H}_K$ associated with the kernel $K(x, x') = \sum_v \beta_v K_v(x, x')$. We define $\mathcal{M} = \sum_v \theta_v \mathcal{M}_v$, where each $\mathcal{M}_v$ is a vector-valued graph Laplacian constructed on $\mathcal{H}_{K_v}$. It can be demonstrated that $\mathcal{M}$ is still a graph Laplacian.

Lemma 1: $\mathcal{M} \in S_{Nn}^{+}$ is a vector-valued graph Laplacian.

The notation S_n^+ denotes a set of $n \times n$ symmetric positive-semidefinite matrices and we will use S_n^* to denote a set of positive-definite matrices. Then we have the following version of the representer theorem.

Theorem 1: For fixed sets of $\{\beta_v\}$ and $\{\theta_v\}$, the minimizer of (6) admits an expansion

$$f^*(x) = \sum_{i=1}^{u+l} K(x, x_i) a_i \tag{7}$$

where $a_i \in \mathcal{Y}, 1 \leq i \leq N = u + l$ are some vectors to be estimated and $K(x, x_i) = \sum_{v=1}^{V} \beta_v K_v(x, x_i)$. The proof of Lemma 1 and Theorem 1 are detailed in the Appendix.

The hinge loss $L(f, x_i, y_i) = (1 - y_i f(x_i))_+$ is more suitable for classification than least squares loss since the hinge loss results in a better convergence rate and usually higher classification accuracy (we refer to [13] for a comparison of different popular loss functions). We adopt the hinge loss in MV^{3}MR and derive MV^{3}LSVM as follows.

C. Multiview Vector-Valued Laplacian SVM

Under the SVM formulation, the minimization problem of MV^{3}MR is

$$\begin{aligned} \underset{f \in \mathcal{H}_K, \beta, \theta}{\operatorname{argmin}} \quad & \frac{1}{nl} \sum_{i=1}^{l} \sum_{j=1}^{n} (1 - y_{ij} f_j(x_i))_+ + \gamma_A \|f\|_K^2 \\ & + \gamma_I \langle \mathbf{f}, \mathcal{M}\mathbf{f} \rangle_{\mathcal{Y}^{u+l}} + \gamma_B \|\beta\|_2^2 + \gamma_C \|\theta\|_2^2 \\ \text{s.t.} \quad & \sum_{v=1}^{V} \beta_v = 1, \quad \beta_v \geq 0, \quad \sum_{v=1}^{V} \theta_v = 1, \quad \theta_v \geq 0 \qquad \forall v. \end{aligned} \tag{8}$$

An unregularized bias b_j is often added to the solution $f_j(x) = \sum_{i=1}^{N} K_j(x, x_i) a_i$ in the SVM formulation. By substituting (7) into the above formulation, we can see the primal problem as follows:

$$\begin{aligned} \underset{\mathbf{a}, b, \xi, \beta, \theta}{\operatorname{argmin}} \quad & \frac{1}{nl} \sum_{i=1}^{l} \sum_{j=1}^{n} \xi_{ij} + \gamma_A \mathbf{a}^T G \mathbf{a} + \gamma_I \mathbf{a}^T G \mathcal{M} G \mathbf{a} \\ & + \gamma_B \|\beta\|_2^2 + \gamma_C \|\theta\|_2^2 \\ \text{s.t.} \quad & y_{ij} \left(\sum_{z=1}^{l+u} K_j(x_i, x_z) a_z + b_j \right) \geq 1 - \xi_{ij}, \xi_{ij} \geq 0 \qquad \forall i, j \\ & \sum_{v=1}^{V} \beta_v = 1, \quad \beta_v \geq 0, \quad \sum_{v=1}^{V} \theta_v = 1, \quad \theta_v \geq 0 \quad \forall v \end{aligned} \tag{9}$$

where $G = \sum_{v=1}^{V} \beta_v G_v$ is the combined vector-valued Gram matrix over the labeled and unlabeled samples defined on kernel K, and $\mathcal{M} = \sum_{v=1}^{V} \theta_v \mathcal{M}_v$ is the integrated vector-valued graph Laplacian. Here, $K_j(\cdot, \cdot)$ is the jth row of the vector-valued kernel K. We have three variables, i.e., $\mathbf{a}$, β, and θ, to be optimized in (9). To solve this problem, we consider the following constrained optimization problem:

$$\begin{aligned} \min \quad & F(\beta, \theta) \\ \text{s.t.} \quad & \sum_{v=1}^{V} \beta_v = 1, \quad \beta_v \geq 0, \quad \sum_{v=1}^{V} \theta_v = 1, \quad \theta_v \geq 0 \quad \forall v \end{aligned} \tag{10}$$

where $F(\beta, \theta)$ equals

$$\begin{cases} \operatorname{argmin}_{\mathbf{a},b,\xi} \frac{1}{nl}\sum_{i=1}^{l}\sum_{j=1}^{n}\xi_{ij}+\gamma_A \mathbf{a}^T G\mathbf{a}+\gamma_I \mathbf{a}^T G\mathcal{M}G\mathbf{a} \\ \quad +\gamma_B\|\beta\|_2^2+\gamma_C\|\theta\|_2^2 \\ \text{s.t.} \quad y_{ij}(\sum_{z=1}^{l+u} K_j(x_i,x_z)a_z+b_j)\geq 1-\xi_{ij}, \\ \quad \xi_{ij}\geq 0, i=1,\ldots,l, j=1,\ldots,n. \end{cases} \tag{11}$$

Here, G and $\mathcal{M}$ take the form as in (9). We can omit the terms $\gamma_B\|\beta\|_2^2$ and $\gamma_C\|\theta\|_2^2$ in (11) since β and θ are fixed. By introducing the Lagrange multipliers μ_{ij} and η_{ij} in (11), we have

$$\begin{aligned} &W(\mathbf{a},\xi,b,\mu,\eta) \\ &=\frac{1}{nl}\sum_{i=1}^{l}\sum_{j=1}^{n}\xi_{ij}+\frac{1}{2}\mathbf{a}^T(2\gamma_A G+2\gamma_I G\mathcal{M}G)\mathbf{a}-\sum_{i=1}^{l}\sum_{j=1}^{n}\eta_{ij}\xi_{ij} \\ &-\sum_{i=1}^{l}\sum_{j=1}^{n}\mu_{ij}\Big(y_{ij}\big(\sum_{z=1}^{l+u}K_j(x_i,x_z)a_z+b_j\big)-1+\xi_{ij}\Big). \end{aligned} \tag{12}$$

By taking the partial derivative w.r.t. ξ_{ij}, b_j, and setting them to be zero, we obtain

$$\begin{aligned} &\frac{\partial W}{\partial b_j}=0\Rightarrow\sum_{i=1}^{l}\mu_{ij}y_{ij}=0, \qquad j=1,\ldots,n \\ &\frac{\partial W}{\partial \xi_{ij}}=0\Rightarrow\frac{1}{nl}-\mu_{ij}-\eta_{ij}=0\Rightarrow 0\leq\mu_{ij}\leq\frac{1}{nl}. \end{aligned}$$

A reduced Lagrangian can be obtained by substituting the above equalities back into (12), which leads to

$$\begin{aligned} W^R(\mathbf{a},\mu)=&\frac{1}{2}\mathbf{a}^T\Big(2\gamma_A G+2\gamma_I G\mathcal{M}G\Big)\mathbf{a}-\mathbf{a}^T G J^T Y_d\mu+\mu^T\mathbf{1} \\ \text{s.t.}\quad &\sum_{i=1}^{l}\mu_{ij}y_{ij}=0, \qquad j=1,\ldots,n, \\ &0\leq\mu_{ij}\leq\frac{1}{nl}, \quad i=1,\ldots,l, \quad j=1,\ldots,n \end{aligned} \tag{13}$$

where $J=[I\ 0]\in\mathbb{R}^{(nl)\times(nl+nu)}$ and I is an $nl\times nl$ identity matrix. Here, $\mu=\{\mu_1,\ldots,\mu_l\}\in\mathbb{R}^{nl}$ is a column vector with each $\mu_i=[\mu_{i1},\ldots,\mu_{in}]^T$, $Y_d=\operatorname{diag}(y_{11},\ldots,y_{1n},\ldots,\ldots,y_{l1},\ldots,y_{ln})$, and $\mathbf{1}$ is an all 1s column vector. Taking the partial derivative of W^R w.r.t. $\mathbf{a}$ and letting it be zero leads to

$$\mathbf{a}^*=(2\gamma_A I+2\gamma_I\mathcal{M}G)^{-1}J^T Y_d\mu^*. \tag{14}$$

Substituting it back into (13) we get

$$\begin{aligned} \mu^*=&\operatorname*{argmax}_{\mu\in\mathbb{R}^{nl}}\mu^T\mathbf{1}-\frac{1}{2}\mu^T S\mu \\ \text{s.t.}\quad &\sum_{i=1}^{l}\mu_{ij}y_{ij}=0, \qquad j=1,\ldots,n \\ &0\leq\mu_{ij}\leq\frac{1}{nl}, \quad i=1,\ldots,l, \quad j=1,\ldots,n \end{aligned} \tag{15}$$

where the matrix $S=Y_d J G(2\gamma_A I+2\gamma_I MG)^{-1}J^T Y_d$. Again, the combined Gram matrix $G=\sum_{v=1}^{V}\beta_v G_v$ and the integrated graph Laplacian $\mathcal{M}=\sum_{v=1}^{V}\theta_v\mathcal{M}_v$. Because of the strong duality, the objective value of (11) is also the objective value of (13), which is $W^R(\mathbf{a}^*,\mu^*)$. Therefore, we can rewrite (10) as

$$\begin{aligned} W(\beta,\theta)=&\ W^R(\mathbf{a}^*,\mu^*)+\gamma_B\|\beta\|_2^2+\gamma_C\|\theta\|_2^2 \\ \text{s.t.}\quad &\sum_{v=1}^{V}\beta_v=1,\beta_v\geq 0;\sum_{v=1}^{V}\theta_v=1,\theta_v\geq 0 \quad \forall v. \end{aligned} \tag{16}$$

For fixed θ, the above problem can be rewritten with respect to β as

$$\begin{aligned} W(\beta)=&\ \beta^T H\beta+\gamma_B\|\beta\|_2^2-h^T\beta \\ \text{s.t.}\quad &\sum_{v}\beta_v=1, \quad \beta\geq 0, \quad v=1,\ldots,V \end{aligned} \tag{17}$$

where $h=[h_1,\ldots,h_V]^T$ with each $h_v=(\mathbf{a}^*)^T G_v J^T Y_d\mu^*-\gamma_A(\mathbf{a}^*)^T G_v\mathbf{a}^*$ and H is a $V\times V$ matrix with the entry $H_{ij}=\gamma_I(\mathbf{a}^*)^T G_i\mathcal{M}G_j\mathbf{a}^*$. We can simply set the derivative of $W(\beta)$ to zero and obtain $\beta=(H+H^T+2\gamma_B I)^{-1}h$. Then the computed β is projected to the positive simplex to satisfy the summation and positive constraints. However, such an approach lacks convergence guarantees and may lead to numerical problems. A coordinate descent algorithm is therefore used to solve (17). In each iteration round during the coordinate descent procedure, two elements β_i and β_j are selected to be updated while the others are fixed. By using the Lagrangian of (17) and considering that $\beta_i+\beta_j$ will not change due to constraint $\sum_{v=1}^{V}\beta_v=1$, we have the following solution for updating β_i and β_j:

$$\begin{cases} \beta_i^*=\frac{2\gamma_B(\beta_i+\beta_j)+(h_i-h_j)+2t_{ij}}{2(H_{ii}-H_{ji}-H_{ij}+H_{jj})+4\gamma_B} \\ \beta_j^*=\beta_i+\beta_j-\beta_i^* \end{cases} \tag{18}$$

where $t_{ij}=(H_{ii}-H_{ji}-H_{ij}+H_{jj})\beta_i-\sum_k(H_{ik}-H_{jk})\beta_k$. The obtained β_i^* or β_j^* may violate the constraint $\beta_v\geq 0$. Thus, we set

$$\begin{aligned} &\beta_i^*=0,\beta_j^*=\beta_i+\beta_j, \text{if } 2\gamma_B(\beta_i+\beta_j)+(h_i-h_j)+2t_{ij}\leq 0 \\ &\beta_j^*=0,\beta_i^*=\beta_i+\beta_j, \text{if } 2\gamma_B(\beta_i+\beta_j)+(h_j-h_i)+2t_{ji}\leq 0. \end{aligned}$$

From (18), we can see that the update criteria tends to assign larger value β_i to larger h_i and smaller H_{ii}. Because $h_i=(\mathbf{a}^*)^T G_i J^T Y_d\mu-\gamma_A(\mathbf{a}^*)^T G_i\mathbf{a}^*$ and $H_{ii}=\gamma_I(\mathbf{a}^*)^T G_i\mathcal{M}G_i\mathbf{a}^*$ measures the discriminative ability and the performance of the ith view. Let $(\mathbf{a}_i^*,\mu_i^*)$ be the solution for the optimization problem of the ith view, which is $W^R(\mathbf{a},\mu)$ with $G=G_i$. If all the solutions are the same, i.e., $(\mathbf{a}_1^*,\mu_1^*)=\cdots=(\mathbf{a}_V^*,\mu_V^*)=(\mathbf{a}^*,\mu^*)$, then the objective value $W^R(\mathbf{a}_i^*,\mu_i^*)$ of the discriminative view tends to be smaller than nondiscriminative view (we assume that all Gram matrices have been normalized). A smaller $W^R(\mathbf{a}_i^*,\mu_i^*)$ corresponds to a larger h_i and a smaller H_{ii}, and thus our algorithm prefers discriminative view. However, the solutions $(\mathbf{a}_1^*,\mu_1^*),\ldots,(a_V^*,\mu_V^*)$ may not be exactly the same as $(\mathbf{a}^*,\mu^*)$. Thus the learned β_i is in general but not strictly consistent with the performance of the ith single view. We can see this in the experiments.

For fixed β, (16) can be simplified as

$$W(\theta)=s^T\theta+\gamma_C\|\theta\|_2^2 \tag{19}$$

$$\text{s.t}\quad \sum_{v}\theta_v=1, \quad \theta_v\geq 0, \quad v=1,\ldots,V \tag{20}$$

Algorithm 1 Optimization Procedure of the Proposed MV^{3}LSVM Algorithm

Input: labeled data $D_l^v = \{(x_i^v, y_i)_{i=1}^l\}$ and unlabeled data $D_u^v = \{(x_i^v)_{i=l+1}^{N=u+l}\}$ form different views, $v = 1, \ldots, V$ is the view index

Algorithm parameters: γ_A, γ_I, γ_B, and γ_C

Output: classifier variable $\mathbf{a}$, the kernel combination coefficients $\{\beta_v\}$, and the graph Laplacian weights $\{\theta_v\}$.

1: Construct the Gram matrix G_v and the vector-valued graph Laplacian $\mathcal{M}_v$ for each view, set $\beta_v = \theta_v = \frac{1}{V}, v = 1, \ldots, V$; compute $G = \sum_{v=1}^V \beta_v G_v$ and $M = \sum_{v=1}^V \theta_v \mathcal{M}_v$.
2: **Iterate**
3: Approximately solve for $\mathbf{a}$ through (14) with fixed G and $\mathcal{M}$
4: Compute β by solving (17) and update the Gram matrix G
5: Compute θ by solving (19) and update the graph Laplacian $\mathcal{M}$
6: **Until convergence**

where $s = [s_1, \ldots, s_V]^T$ with each $s_v = \gamma_I(\mathbf{a}^*)^T G\mathcal{M}_v G\mathbf{a}^*$. Similarly, the solution of (19) can be obtained by using the coordinate descent, and the criteria for updating θ_i and θ_j in an iteration round is given by

$$\begin{cases} \theta_i^* = 0, \theta_j^* = \theta_i + \theta_j, & \text{if } 2\gamma_C(\theta_i + \theta_j) + (s_j - s_i) \le 0 \\ \theta_j^* = 0, \theta_i^* = \theta_i + \theta_j, & \text{if } 2\gamma_C(\theta_i + \theta_j) + (s_i - s_j) \le 0 \\ \theta_i^* = \frac{2\gamma_C(\theta_i+\theta_j)+(s_j-s_i)}{4\gamma_C}, \theta_j^* = \theta_i + \theta_j - \theta_i^*, & \text{else.} \end{cases} \quad (21)$$

We now summarize the learning procedure of the proposed MV^{3}LSVM in Algorithm 1.

The stopping criterion for terminating the algorithm can be the difference of the objective value $W^R(\mathbf{a}, \mu) + \gamma_B\|\beta\|_2^2 + \gamma_C\|\theta\|_2^2$ between two consecutive steps. Alternatively, we can stop the iterations when the variation of β and θ are both smaller than a predefined threshold. Our implementation is based on the difference of the objective value, i.e., if the value $|O_k - O_{k-1}|/|O_k - O_0|$ is smaller than a predefined threshold, then the iteration stops, where O_k is the objective value of the kth iteration step. Our implementation is based on the difference of the objective value.

D. Convergence Analysis

In this section, we discuss the convergence of the proposed MV^{3}LSVM algorithm. We first prove the convexity of (11), (17), and (19) as follows.

Proof: The Hessian matrix of the objective function of (11) is $H_e(\mathbf{a}) = \gamma_A G + \gamma_I G\mathcal{M}G$. The Gram matrix $G \in S_n^+$ and we assume that G is positive definite in this paper (to enforce this property a small ridge is added to the diagonal of G). The second term is positive semidefinite since $x^T G\mathcal{M}Gx = z^T\mathcal{M}z \ge 0$ for any x and $z = Gx$. Here, we have used the property of the graph Laplacian $\mathcal{M} \in S_{Nn}^+$. Then $H_e(\mathbf{a}) \in S_{Nn}^*$ for $\gamma_A > 0$ and (11) is strictly convex.

For (17), the Hessian matrix is $H_e(\beta) = H + \gamma_B I$. The matrix H is symmetric since the element $H_{ij} = H_{ij}^T = \gamma_I \mathbf{a}^T G_j \mathcal{M} G_i \mathbf{a} = H_{ji}$. In addition, the Cholesky decomposition $\mathcal{M} = P^T P$ exists since $\mathcal{M} \in S_{Nn}^+$. Let $z_i = PG_i a$, and we have $H_{ij} = \gamma_I z_i^T z_j$. Thus, $H \in S_V^+$ and $H_e(\beta) \in S_V^*$ for $\gamma_B > 0$. This means that (17) is also strictly convex.

Finally, it is straightforward to verify that (19) is strictly convex for $\gamma_C > 0$. This completes the proof. ■

Now we discuss the convergence of our algorithm. Let the objective function of (9) be $R(\mathbf{a}, b, \xi, \beta, \theta)$ and the initialized value be $R(\mathbf{a}^k, b^k, \xi^k, \beta^k, \theta^k)$. Since (11) is convex, we have $R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^k, \theta^k) \le R(\mathbf{a}^k, b^k, \xi^k, \beta^k, \theta^k)$. We suppose that (9) is exactly solved, which means that the duality gap is zero. Then $R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^k, \theta^k) = W(\beta^k, \theta^k)$. For fixed θ^k, we obtain the convex problem (17), thus we have $R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^{k+1}, \theta^k) \le R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^k, \theta^k)$. Similarly, due to the convexity of (19), we have $R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^{k+1}, \theta^{k+1}) \le R(\mathbf{a}^{k+1}, b^{k+1}, \xi^{k+1}, \beta^{k+1}, \theta^k)$. Therefore, the convergence of our algorithm is guaranteed.

E. Complexity Analysis

For the proposed MV^{3}LSVM, the complexity is dominated by the time cost of computing a^* in each iteration, where the computation of the matrix S in (15) involves an inversion and several multiplications of $nN \times nN$ matrix, and the time complexity is $O(n^{2.8}N^{2.8})$ using the Strassen algorithm [53]. (15) can be solved using a standard SVM solver with the time complexity $O(n^{2.3}l^{2.3})$ according to the sequential minimal optimization [54]. The computations of β and θ are quite efficient since their dimensionality is V, which is usually very small (e.g., $V = 7$ in our experiments). Suppose the number of iterations is k, then the total cost of MV^{3}LSVM is $O(k(n^{2.8}N^{2.8} + n^{2.3}l^{2.3}))$. Considering that $l < N$, the time cost is $O(kn^{2.8}N^{2.8})$, which is k times of the case where no combination coefficients (β and θ) are learned. From the experimental results shown in Section V-B, we will find that k is very small since our algorithm only needs a few iterations (around five) to converge. Actually, there is a balance between the time complexity and classification accuracy. If only a limited number of unlabeled samples are selected to construct the input graph Laplacians, i.e., $N = u+l$ is small, then the time complexity can be reduced with an acceptable performance sacrifice. In our experiments, we obtain satisfactory accuracy by setting $N = 1000$, so the time cost is acceptable.

V. Experiment

We validate the effectiveness of MV^{3}LSVM on two challenge datasets, VOC [16] and MIR [17]. The VOC dataset contains 10 000 images labeled with 20 categories. The MIR dataset consists of 25 000 images of 38 concepts. For the VOC dataset [16], we use the standard train/test partition [16], which splits 9963 images into a training set of 5011 images and a test set of 4952 images. For the MIR dataset [17], images are randomly split into equally sized training and test sets. For both datasets, we randomly select 20% of the test images for validation and the rest for testing. The parameters of all the algorithms compared in our experiments are tuned by using the validation set. This means that the parameters corresponding

to the best performance in the validation set are used for the transductive inference and inductive test. From the training examples, 10 random choices of $l \in \{100, 200, 500\}$ labeled samples are used in our experiments.

We use several visual views and the tag feature according to [2]. The visual views include SIFT features [3], local hue histograms [55], global GIST descriptor [4], and some color histograms (RGB, HSV, and LAB). The local descriptors (SIFT and hue) are computed densely on the multiscale grid and quantized using k-means, which results in a visual word histogram for each image. Therefore, we have seven different representations in total. We precompute a scalar-valued Gram matrix for each view and normalize it to unit trace. For the visual representations, the kernel is defined by

$$k(x_i, x_j) = \exp\big(-\lambda^{-1} d(x_i, x_j)\big)$$

where $d(x_i, x_j)$ denotes the distance between x_i and x_j, and $\lambda = \max_{i,j} d(x_i, x_j)$. Following [2], we choose the $L1$ distance for the color histogram representations (RGB, HSV, and LAB), and $L2$ for GIST and χ^2 for the visual word histograms (SIFT and hue). For the tag features, a linear kernel $k(x_i, x_j) = x_i^T x_j$ is constructed.

A. Evaluation Metrics

We use three kinds of evaluation criteria. The AP and AUC are utilized to evaluate the ranking performance under each label. We also use the RL to study the performance of label set prediction for each instance.

1) AP evaluates the fraction of samples ranked above a particular positive sample [56]. For each label, there is a ranked sequence of samples returned by the classifier. A good classifier will rank most of the positive samples higher than the negative ones. The traditional AP is defined as

$$\mathrm{AP} = \frac{\sum_k \mathrm{P}(k)}{\#\{\text{positive samples}\}}$$

where k is a rank index of a positive sample and $\mathrm{P}(k)$ is the precision at the cutoff k. In this paper, we choose to use the computing method as in the PASCAL VOC [16] challenge evaluation

$$\mathrm{AP} = \frac{1}{11} \sum_r \mathrm{P}(r)$$

where $\mathrm{P}(r)$ is the maximum precision over all recalls larger than $r \in \{0, 0.1, 0.2, \ldots, 1.0\}$. A larger value means a higher performance. In this paper, the mean AP, i.e., mAP, over all labels, is reported to save space.

2) ROC evaluates the probability that a positive sample will be ranked higher than a negative one by a classifier [57]. It is computed from an ROC curve that depicts relative tradeoffs between true positive (benefits) and false positive (costs). The AUC of a realistic classifier should be larger than 0.5. We refer to [57] for a detailed description. A larger value means a higher performance. Similar to AP, the mean AUC, i.e., mAUC, over all labels, is reported.

3) RL evaluates the fraction of label pairs that are incorrectly ranked [24], [26]. Given a sample x_i and its label set Y_i, a successful classifier $f(x, y)$ should have a larger value for $y \in Y_i$ than those $y \notin Y_i$. Then the RL for the ith sample is defined as

$$\mathrm{RL}(f, x_i) = \frac{1}{|Y_i|(P - |Y_i|)} \times |\{(y_1, y_2) | f(x, y_1) \leq f(x, y_2), y_1 \in Y_i, y_2 \notin Y_i\}|$$

where P is the total number of labels and $|\cdot|$ denotes the cardinality of a set. The smaller the value, the higher the performance. The mean value over all samples is computed for evaluation.

B. Performance Enhancement With Multiview Learning

It has been shown in [8] that VVMR performs well for transductive semisupervised multilabel classification and can provide a high-quality out-of-sample generalization. The proposed MV^{3}MR framework is a multiview generalization of VVMR that incorporates the advantage from MKL for handling multiview data. Therefore, we first evaluate the effectiveness of learning the view combination weights using the proposed multiview learning algorithm for transductive semisupervised multilabel classification. An out-of-sample evaluation will be presented in the next subsection. The experimental setup of the two compared methods is given as follows.

1) *VVLSVM:* Vector-valued LSVM is an SVM implementation of the vector-valued manifold regularization framework that exploits both the geometry of the input data as well as the label correlations. We do not use the vector-valued Laplacian RLS presented in [8] for comparison because the hinge loss is more suitable for classification. The parameters γ_A and γ_I in (2) are both optimized over the set $\{10^i | i = -8, -7, \ldots, -2, -1\}$. We set the parameter γ_O in (3) to 1.0 since it has been demonstrated empirically in [8] that with a larger γ_O, the performance will usually be better. The mean of the multiple Gram matrices and input graph Laplacians are precomputed for the experiments. The number of nearest neighbors for constructing the input and output graph Laplacians are tuned on the sets $\{10, 20, \ldots, 100\}$ and $\{2, 4, \ldots, 20\}$, respectively.
2) *MV^{3}LSVM:* This an SVM implementation of the proposed MV^{3}MR framework that combines multiple views by constructing kernels for all views and learning their weights. We tune the parameters γ_A and γ_I as in VVLSVM and γ_O is set to 1.0. The additional parameters γ_B and γ_C are optimized over $\{10^i | i = -8, -7, \ldots, -2, -1\}$. We first only learn kernel combinations β and set the graph weights θ to be uniform (MV^{3}LSVM1 in Fig. 4). Then we learn both and θ in MV^{3}LSVM2. We use 20 and 6 nearest neighbor graphs to construct the input and output normalized graph Laplacians, respectively, for the VOC dataset, and 30 and 8 nearest neighbor graphs in the experiments on MIR. We set these hyperparameters to be the same as those in VVLSVM and no further optimization was attempted.

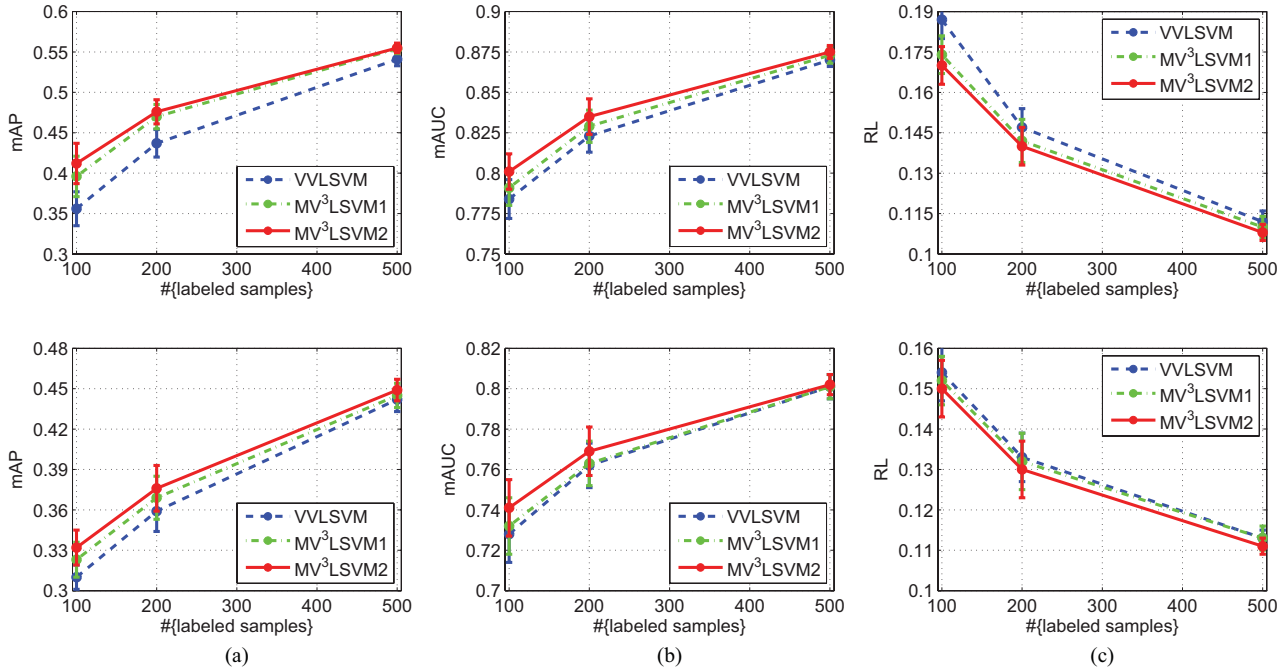


Fig. 4. (a) mAP, (b) mAUC, and (c) RL performance enhancement by learning the weights (β and θ) for different views. PASCAL VOC' 07 (top) and MIR Flickr (bottom). MV3LSVM1: only learn β; MV3LSVM2: learn both β and θ.

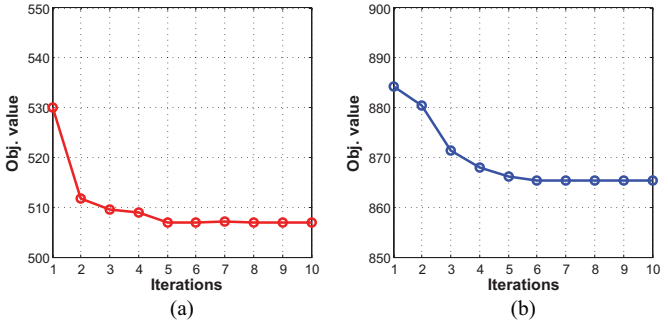


Fig. 5. Behavior of the objective values by increasing the iteration number on the two datasets. (a) PASCAL VOC' 07. (b) MIR Flickr.

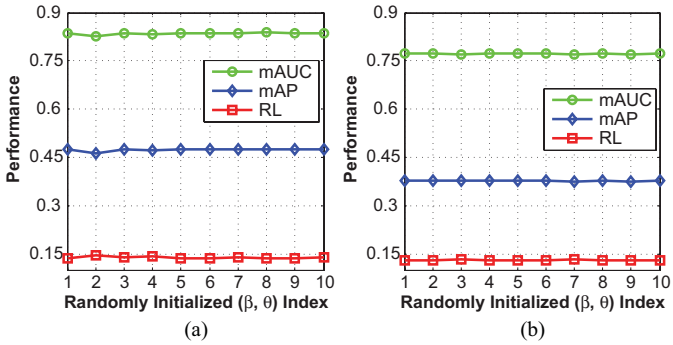


Fig. 6. Performance in mAP, mAUC, and RL verses randomly initialized (β, θ). The proposed model is insensitive to different initializations of (β, θ). (a) PASCAL VOC' 07. (b) MIR Flickr.

The experimental results on the two datasets are shown in Fig. 4. We can see that learning the combination weights using our algorithm is always superior to simply using the uniform weights for different views. We also find that, when the number of labeled samples increases, the improvement becomes small. This is because the multiview learning actually helps to approximate the underlying data distribution. This approximation can be steadily improved with the increase of the number of labeled samples, and thus the significance of the multiview learning to the approximation gradually decreases. Besides, we observe that β has more influence on the final performance overall.

We show the behavior of the objective values by increasing the iteration number in Fig. 5. From the figure, we can see that only a few iterations (about five) are necessary to obtain a satisfactory solution. Thus the time complexity is only a little more than the VVLSVM algorithm and can justify the performance enhancement.

Finally, our algorithm is not sensitive to different initializations, as shown in Fig. 6. In particular, we run our algorithm with 10 random choices of β and θ. We show the performance in terms of mAP, mAUC, and RL on the two datasets in Fig. 6. It can be observed that the performance curves do not vary a lot with different initializations.

C. Out-of-Sample Generalization

The second set of experiments is to evaluate the out-of-sample extension quality of the MV^{3}MR framework, and the SVM implementation is utilized. Fig. 7 compares the transductive performance to the inductive performance when using $l = 200$ labeled samples. We show a scatter plot of the AP scores for each label on the two datasets by using 10 random choices of labeled data. We can see that our algorithm generalizes well from the unlabeled set to the unseen set. The MV^{3}MR framework inherits a strong natural out-of-sample generalization ability that many semisupervised multilabel methods do not naturally have [8]. Besides, most graph-based semisupervised learning algorithms are transductive and

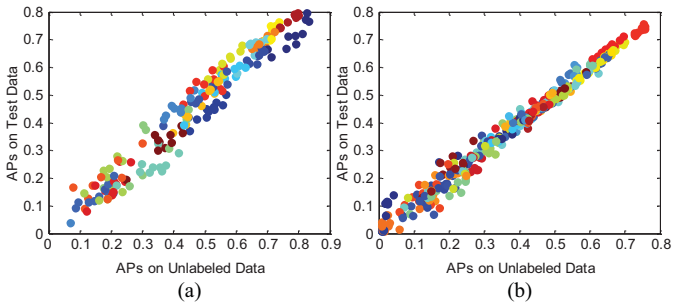


Fig. 7. Transductive and inductive APs across outputs on (a) VOC and (b) MIR datasets.

additional induction schemes are necessary to handle new points [58].

D. Analysis of the Combination Coefficients in Multiview Learning

In the following, we present empirical analyses of the multiview learning procedure. In Fig. 8, we select $l = 200$ and present the view combination coefficients β and θ learned by MV^{3}LSVM, together with the mAP by using VVLSVM for each view. From the results, we find that the tendency of the kernel and graph weights are both consistent with the corresponding mAP in general, i.e., the views with a higher classification performance tend to be assigned larger weights, taking the DenseSIFT visual view (the second view) and the tag (the last view) for example. However, a larger weight may sometimes be assigned to a less discriminative view; for example, the weight of Hsv (the fourth view) is larger than the weight of DenseSift (the second view). This is mainly because the coefficient $\mathbf{a}$ is not optimal for every single view, in which only G_v and $\mathcal{M}_v$ are utilized. The learned $\mathbf{a}$ minimizes the optimization problem (8) by using the combined Gram matrix G and integrated graph Laplacian $\mathcal{M}$, which means that the learned vector-valued function is smooth along the combined RKHS and the integrated manifold. In this way, the proposed algorithm effectively exploits the complementary property of different views.

E. Comparisons With Multilabel and MKL Algorithms

Our last set of experiments is to compare MV^{3}LSVM with several competitive multilabel methods as well as with some well-known and competitive MKL algorithms in predicting the unknown labels of the unlabeled data. The out-of-sample generalization ability of our method has been verified in our second set of experiments.

We specifically compare MV^{3}LSVM with the following methods on the challenging VOC and MIR datasets

1) *SVM_CAT:* Concatenating the features of each view and then running standard SVM. The parameter C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$. The time complexity is $O(nl^{2.3})$ [54].
2) *SVM_UNI:* Combining different kernels by combining them with uniform weights and then running standard SVM. The parameter C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$. The time complexity is $O(nl^{2.3})$ [54].
3) *MLCS [20]:* MLCS algorithm that takes advantage of the sparsity of the labels. We choose the label compression ratio to be 1.0 since the number of the labels n is not very large. The mean of the multiple kernels from different views is precomputed for the experiments. Suppose the length of the compressed label vector (for each sample) is $r \leq n$. Then the training cost is $O(nl^3)$ if we choose the regression algorithm to be the least squares [59], and the reconstruction complexity is $O(l(n^3 + rn^2))$ if we use the least angle regression algorithm [60]. Considering that $r \leq n \leq l$ in this paper, the time complexity of MLCS is $O(nl^3)$.
4) *KLS_CCA [12]:* A least-squares formulation of the kernelized CCA for multilabel classification. The ridge parameter is chosen from the candidate set $\{0, 10^i | i = -3, -2, \ldots, 2, 3\}$. The mean of multiple Gram matrices is precomputed to run the algorithm. According to the discussion presented in [12], the time complexity is $O(n^2l + kn(3l + 5d + 2dl))$, where d is the feature dimensionality and k is the number of iterations.
5) *SimpleMKL [18]:* A popular SVM-based MKL algorithm that determines the combination of multiple kernels by a reduced gradient descent algorithm. The penalty factor C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$. We apply SimpleMKL to multilabel classification by learning a binary classifier for each label. According to the Algorithm 1 presented in [18], there is an outer loop for updating the kernel weights, as well as an inner loop to determine the maximal admissible step size in the reduced gradient descent. Suppose the number of outer and inner iterations are k_1 and k_2, respectively; then the time complexity of SimpleMKL is approximately $O(nk_1k_2l^{2.3})$, where we have ignored the time cost of the SVM solver in the inner loop since it has warm start and can be very fast [18].
6) *LpMKL [19]:* A recent proposed MKL algorithm, which extend MKL to l_p-norm with $p \geq 1$. The penalty factor C is tuned on the set $\{10^i | i = -1, 0, \ldots, 7, 8\}$ and we choose the norm p from the set $\{1, 8/7, 4/3, 2, 4, 8, 16, \infty\}$. According to Algorithm 1 presented in [19], the time complexity is $O(nkl^{2.3})$ since the kernel combination coefficients can be computed analytically, where k is the number of iterations.

The performance of the compared methods on the VOC dataset and MIR dataset are reported in Table I. The values in the last column of Table I are the average ranks. From the results, we first observe that the performance keeps improving with increasing number of the labeled samples. Second, the performance of the simpleMKL algorithm, which learns the kernel weights for SVM, can be inferior to that of the multilabel algorithms with the mean kernel in many cases. MV^{3}LSVM is superior to multiview (SimpleMKL and LpMKL) and multilabel algorithms in general and consistently outperforms other methods in terms of mAP. The average rank of our algorithm is smaller than those of all the other methods in terms of all the three criteria. According to the Friedman test

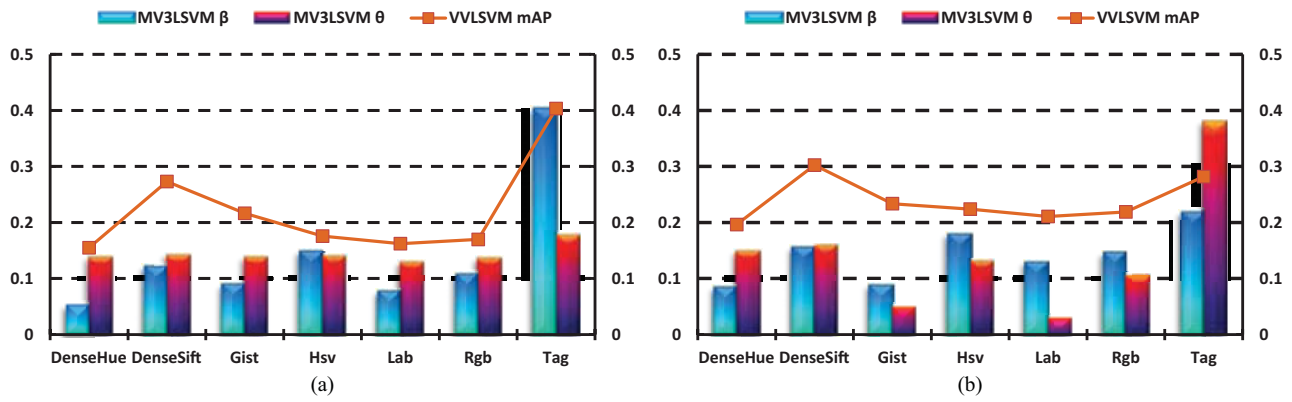


Fig. 8. View combination weights β and θ learned by MV^{3}MR, as well as the mAP of using VVLSVM for each view. (a) PASCAL VOC' 07. (b) MIR Flickr.

TABLE I
PERFORMANCE EVALUATION ON THE TWO DATASETS

	VOC mAP ↑ Versus #{Labeled Samples}			MIR mAP ↑ Versus #{Labeled Samples}			
Methods	100	200	500	100	200	500	Ranks
SVM_CAT	0.241±0.011 (7)	0.288±0.013 (7)	0.371±0.007 (7)	0.281±0.009 (7)	0.306±0.007 (7)	0.352±0.008 (7)	7
SVM_UNI	0.347±0.018 (4.5)	0.424±0.014 (5)	0.529±0.006 (5)	0.302±0.011 (5)	0.336±0.013 (6)	0.400±0.009 (6)	5.25
MLCS [20]	0.332±0.017 (6)	0.412±0.016 (6)	0.525±0.007 (6)	0.289±0.010 (6)	0.342±0.011 (5)	0.424±0.010 (5)	5.67
KLS_CCA [12]	0.347±0.019 (4.5)	0.432±0.014 (4)	0.536±0.007 (4)	0.321±0.009 (3.5)	0.369±0.017 (2)	0.445±0.009 (2.5)	3.42
MV^{3}LSVM	**0.412±0.025 (1)**	**0.476±0.015 (1)**	**0.555±0.006 (1)**	**0.332±0.013 (1)**	**0.376±0.017 (1)**	**0.449±0.008 (1)**	**1**
SimpleMKL [18]	0.381±0.024 (3)	0.453±0.020 (3)	0.538±0.011 (3)	0.321±0.014 (3.5)	0.365±0.017 (4)	0.444±0.011 (2.5)	3.17
LpMKL [19]	0.391±0.024 (2)	0.462±0.012 (2)	0.540±0.006 (2)	0.327±0.010 (2)	0.367±0.014 (3)	0.436±0.008 (4)	2.5
	VOC mAUC ↑ Versus #{Labeled Samples}			MIR mAUC ↑ Versus #{Labeled Samples}			
Methods	100	200	500	100	200	500	Ranks
SVM_CAT	0.744±0.013 (7)	0.785±0.006 (7)	0.832±0.003 (7)	0.722±0.008 (4)	0.745±0.004 (6)	0.783±0.004 (6)	6.17
SVM_UNI	0.783±0.008 (3)	0.824±0.009 (2.5)	0.870±0.003 (2.5)	0.718±0.011 (5)	0.742±0.011 (7)	0.782±0.006 (7)	4.5
MLCS [20]	0.773±0.010 (5)	0.819±0.010 (6)	0.869±0.004 (4)	0.701±0.012 (7)	0.749±0.010 (5)	**0.805±0.005 (1.5)**	4.42
KLS_CCA [12]	0.781±0.009 (4)	0.824±0.008 (2.5)	0.866±0.003 (5)	0.737±0.009 (2)	**0.769±0.010 (1.5)**	**0.805±0.005 (1.5)**	2.75
MV^{3}LSVM	**0.801±0.011 (1)**	**0.835±0.011 (1)**	**0.875±0.004 (1)**	**0.741±0.014 (1)**	**0.769±0.012 (1.5)**	0.802±0.005 (3.5)	**1.5**
SimpleMKL [18]	0.769±0.017 (6)	0.822±0.013 (4.5)	0.870±0.006 (2.5)	0.717±0.013 (6)	0.753±0.010 (4)	0.802±0.005 (3.5)	4.42
LpMKL [19]	0.786±0.008 (2)	0.822±0.008 (4.5)	0.862±0.005 (6)	0.732±0.010 (3)	0.756±0.010 (3)	0.795±0.007 (5)	3.92
	VOC RL ↓ Versus #{Labeled Samples}			MIR RL ↓ Versus #{Labeled Samples}			
Methods	100	200	500	100	200	500	Ranks
SVM_CAT	0.220±0.008 (7)	0.183±0.006 (7)	0.142±0.003 (7)	0.165±0.005 (3)	0.146±0.004 (5)	0.126±0.002 (5)	5.67
SVM_UNI	0.178±0.008 (2)	0.143±0.006 (3)	**0.106±0.003 (1.5)**	0.549±0.040 (7)	0.437±0.022 (7)	0.177±0.011 (7)	4.58
MLCS [20]	0.195±0.007 (5)	0.155±0.007 (5)	0.112±0.004 (4)	0.173±0.006 (5)	0.145±0.005 (4)	0.115±0.003 (2.5)	4.25
KLS_CCA [12]	0.183±0.008 (3)	0.149±0.006 (4)	0.122±0.005 (5)	0.168±0.013 (4)	0.143±0.005 (3)	0.121±0.004 (4)	3.83
MV^{3}LSVM	**0.170±0.007 (1)**	**0.140±0.007 (1)**	0.108±0.003 (3)	**0.150±0.006 (1)**	**0.130±0.007 (1)**	**0.111±0.003 (1)**	**1.33**
SimpleMKL [18]	0.214±0.017 (6)	0.142±0.009 (2)	**0.106±0.004 (1.5)**	0.155±0.005 (2)	0.136±0.006 (2)	0.115±0.003 (2.5)	2.67
LpMKL [19]	0.186±0.011 (4)	0.164±0.010 (6)	0.137±0.007 (6)	0.199±0.014 (6)	0.181±0.007 (6)	0.141±0.004 (6)	5.67

(↑ indicates "the larger the better"; ↓ indicates "the smaller the better." Mean and std. are reported. The best result is highlighted in boldface.)

[61], the statistics F_F of mAP, mAUC and RL are 56.05, 3.03, and 5.69 respectively. All of them are larger than the critical value $F(6, 30) = 2.42$, so we reject the null hypothesis (the compared algorithms perform equally well). In particular, by comparing it with SimpleMKL, we obtain a significant 8.1% mAP improvement on VOC when using 100 labeled samples. The level of improvement drops when more labeled samples are available, for the same reason described in our first set of experiments.

VI. CONCLUSION

Most of the existing works on multilabel image classification use only single feature representation, and the multiple

feature methods usually assume that a single label is assigned to an image. However, an image is usually associated with multiple labels and different kinds of features are necessary to describe the image properly. Therefore, we have developed a MV^{3}MR for multilabel image classification in which images are naturally characterized by multiple views. MV^{3}MR combines different kinds of features in the learning process of the vector-valued function for multilabel classification. We also derived an SVM formulation of MV^{3}MR, which results in MV^{3}LSVM. The new algorithm effectively exploits the label correlations and learns the view weights to integrate the consistency and complementary properties of different views. We evaluated the proposed algorithm in terms of three popular criteria, i.e., mAP, mAUC, and RL. Intensive experiments on two challenge datasets VOC and MIR showed that the SVM-based implementation under MV^{3}MR outperforms the traditional multilabel algorithms as well as some well-known multiple kernel learning methods. Furthermore, our method provides a strategy for learning from multiple views in multilabel classification and can be extended to other multilabel algorithms.

APPENDIX A
PROOF OF LEMMA 1

Proof: The matrix $\mathcal{M} = \sum_v \theta_v \mathcal{M}_v = \sum_v \theta_v (\mathcal{L}_v \otimes I_n) = \mathcal{L} \otimes I_n$, where $\mathcal{L} = \sum_v \theta_v \mathcal{L}_v$ is defined as a convex combination of the scalar-valued graph Laplacians constructed from different views. $\mathcal{L} \in S_N^+$ since each $\mathcal{L}_v \in S_N^+$, and thus we have $\mathcal{M} \in S_{Nn}^+$ according to the positive-definite property on the Kronecker product. Here, $\mathcal{L} = \sum_v \theta_v (\mathcal{D}_v - \mathcal{W}_v)$ can be computed by using the following adjacency graph:

$$\mathcal{W}_{ij} = \begin{cases} \sum_v \theta_v \mathcal{W}_{vij}, & \text{if } x_i \in N(x_j) \text{ or } x_j \in N(x_i), \\ 0, & \text{otherwise} \end{cases}$$

where $N(x)$ denotes a set that contains the KNNs of x, and $\mathcal{W}_{vij}$ is the similarity between the ith and jth point from the vth view. Thus $\mathcal{L}$ is a graph Laplacian and $\mathcal{M}$ is the corresponding vector-valued graph Laplacian. ■

APPENDIX B
PROOF OF THE REPRESENTER THEOREM

Proof: It has been presented in Section IV-A that there is an RKHS $\mathcal{H}_K$ associated with the vector-valued kernel K. The probability distribution is assumed to be supported on a manifold M in the manifold regularization framework. We now denote $S = \{\sum_i K(x_i, \cdot) a_i | x_i \in M, a_i \in \mathcal{Y}\}$ as a linear space spanned by the kernels centered at the points on M. Any function $f \in \mathcal{H}_K$ can be decomposed as $f = f_\| + f_\perp$, with $f_\| \in S$ and $f_\perp \in S^\perp$. It has been proved in Lemma 1 that $\mathcal{M}$ is a graph Laplacian. Thus we can use $\mathcal{M}$ to induce an intrinsic norm $\|\cdot\|_I$, which satisfies $\|f\|_I = \|g\|_I$ for any $f, g \in \mathcal{H}_K$, $(f - g)|_M \equiv 0$. According to the reproducing property, it concludes that $f_\perp$ vanishes on M [9]. This means that for any $x \in M$, we have $f(x) = f_\|(x)$ and then $\|f\|_I = \|f_\|\|_I$. Besides $\|f\|_K^2 = \|f_\|\|_K^2 + \|f_\perp\|_K^2 \geq \|f_\|\|_K^2$, and thus we conclude that the minimizer of (6) must lie in S for fixed β and θ. Furthermore, because M is approximated by the Laplacian of the graph constructed by the labeled and unlabeled samples, we have $S = \{\sum_{i=1}^{l+u} K(x_i, \cdot) a_i | a_i \in \mathcal{Y}\}$. This completes the proof. ■

ACKNOWLEDGMENT

The authors would like to thank the Editor-In-Chief Prof. Dr. D. Liu, the handling associate editor and all four anonymous reviewers for their positive support and constructive comments to this paper. The authors would also like to thank Journals Coordinator M. Hellrigel for careful editing.

REFERENCES

[1] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.
[2] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 902–909.
[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[5] E. Çesmeli and D. Wang, "Texture segmentation using gaussian-markov random fields and neural oscillator networks," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 394–404, Mar. 2001.
[6] D. Masip and J. Vitrià, "Shared feature extraction for nearest neighbor face recognition," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 586–595, Apr. 2008.
[7] C. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Comput.*, vol. 17, no. 1, pp. 177–204, 2005.
[8] H. Minh and V. Sindhwani, "Vector-valued manifold regularization," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 57–64.
[9] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 2399–2434, Jan. 2006.
[10] G. Chen, Y. Song, F. Wang, and C. Zhang, "Semi-supervised multi-label learning by solving a sylvester equation," in *Proc. SIAM Int. Conf. Data Mining*, 2008, pp. 410–419.
[11] B. Hariharan, L. Zelnik-Manor, S. Vishwanathan, and M. Varma, "Large scale max-margin multi-label classification with priors," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 423–430.
[12] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.
[13] L. Rosasco, E. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?" *Neural Comput.*, vol. 16, no. 5, pp. 1063–1076, 2004.
[14] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1088–1099, Jul. 2006.
[15] D. Tao, X. Li, and S. Maybank, "Negative samples analysis in relevance feedback," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 4, pp. 568–580, Apr. 2007.
[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge (VOC) results," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[17] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, 2008, pp. 39–43.
[18] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.
[19] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "Lp-norm multiple kernel learning," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 953–997, Mar. 2011.
[20] D. Hsu, S. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," in *Advances in Neural Information Processing Systems*. New York, USA: Springer-Verlag, 2009, pp. 772–780.
[21] T. Zhou and D. Tao, "Multi-label subspace ensemble," *J. Mach. Learn. Res.*, vol. 22, pp. 1444–1452, Apr. 2012.

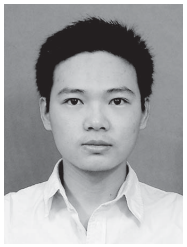
[22] T. Zhou, D. Tao, and X. Wu, "Compressed labeling on distilled labelsets for multi-label learning," *Mach. Learn.*, vol. 88, no. 1, pp. 69–126, 2012.
[23] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multi-task learning for semi-supervised multi-label image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.
[24] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, no. 2, pp. 135–168, 2000.
[25] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in Neural Information Processing Systems*. New York, USA: Springer-Verlag, 2001, pp. 681–687.
[26] M. Zhang and Z. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.
[27] A. McCallum, "Multi-label text classification with a mixture model trained by EM," in *Proc. Assoc. Adv. Artif. Intell. Workshop Text Learn.*, 1999, pp. 1–7.
[28] L. Sun, S. Ji, and J. Ye, "Hypergraph spectral learning for multi-label classification," in *Proc. 14th ACM Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 668–676.
[29] S. Ji, L. Tang, S. Yu, and J. Ye, "A shared-subspace learning framework for multi-label classification," *ACM Trans. Knowl. Discovery Data*, vol. 4, no. 2, pp. 1–29, 2010.
[30] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.
[31] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.
[32] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," in *Proc. 19th Int. Conf. Mach. Learn.*, Jan. 2002, pp. 323–330.
[33] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, Jul. 2006.
[34] K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, no. 2, pp. 349–396, 2003.
[35] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2006, pp. 2169–2178.
[36] K. Labusch, E. Barth, and T. Martinetz, "Simple method for high-performance digit recognition based on sparse coding," *IEEE Trans. Neural Netw.*, vol. 19, no. 11, pp. 1985–1989, Nov. 2008.
[37] X. Zhou, K. Yu, T. Zhang, and T. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 141–154.
[38] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1465–1472.
[39] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.
[40] B. Xie, Y. Mu, D. Tao, and K. Huang, "m-SNE: Multiview stochastic neighbor embedding," *IEEE Trans. Systems, Man, Cybern., B, Cybern.*, vol. 41, no. 4, pp. 1088–1096, Aug. 2011.
[41] J. Yu, M. Wang, and D. Tao, "Semi-supervised multiview distance metric learning for cartoon synthesis," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4636–4648, Nov. 2012.
[42] H. Su, M. Sun, L. Fei-Fei, and S. Savarese, "Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 213–220.
[43] Y. Fu, Y. Guo, Y. Zhu, F. Liu, C. Song, and Z. Zhou, "Multi-view video summarization," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 717–729, Aug. 2010.
[44] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial Neural Networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 412–424, Jul. 2012.
[45] A. Saffari, C. Leistner, M. Godec, and H. Bischof, "Robust multi-view boosting with priors," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 776–789.
[46] D. Zhang, F. Wang, L. Si, and T. Li, "Maximum margin multiple instance clustering with its applications to image and text clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 739–751, May 2011.
[47] H. Zhang, G. Liu, T. Chow, and W. Liu, "Textual and visual content-based anti-phishing: A Bayesian approach," *IEEE Trans. Neural Netw.*, vol. 22, no. 10, pp. 1532–1546, Oct. 2011.
[48] S. Bucak, R. Jin, and A. Jain, "Multi-label learning with incomplete class assignments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2801–2808.
[49] X. Chen, X. Yuan, Q. Chen, S. Yan, and T. Chua, "Multi-label visual classification with label exclusive context," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 834–841.
[50] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Nural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.
[51] D. Bouchaffra, "Mapping dynamic bayesian networks to α-shapes: Application to human faces identification across ages," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1229–1241, Aug. 2012.
[52] L. Chen, I. Tsang, and D. Xu, "Laplacian embedded regression for scalable manifold regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 902–915, Jun. 2012.
[53] V. Strassen, "Gaussian elimination is not optimal," *Numer. Math.*, vol. 13, no. 4, pp. 354–356, 1969.
[54] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*. Cambridge, MA, USA: MIT Press, pp. 185–208, 1999.
[55] J. Van De Weijer and C. Schmid, "Coloring local feature extraction," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 334–348.
[56] M. Zhu, "Recall, precision and average precision," Dept. Electr. Eng., Univ. Waterloo, Waterloo, ON, Canada, Tech. Rep. 2004-09, 2004.
[57] T. Fawcett, "Roc graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, pp. 1–38, Mar. 2004.
[58] X. Zhu, "Semi-supervised learning literature survey," Dept. Electr. Eng., Univ. Wisconsin Madison, Madison, WI, USA, Tech. Rep. 1530, 2006.
[59] M. Gori, "On the time complexity of regularized least square," in *Proc. WIRN*, 2011, pp. 85–96.
[60] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals Stat.*, vol. 32, no. 2, pp. 407–499, 2004.
[61] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, Jan. 2006.

Yong Luo received the B.Sc. degree from Northwestern Polytechnical University, Xi'an, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China.

He was a Visiting Student with the School of Computing, Nanyang Technological University, Singapore, and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia. His current research interests include machine learning and its applications on image classification and annotation.

Dacheng Tao (M'07–SM'12) is currently a Professor of computer science with the Centre for Quantum Computation and Information Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia. He has authored or co-authored more than 100 papers in journals and conferences at top venues including the IEEE T-PAMI, T-IP, T-NNLS, CVPR, ECCV, and ICDM. His current research interests include applying statistics and mathematics for data analysis problems in data mining, computer vision, machine learning, multimedia, and video surveillance.

Dr. Tao was a recipient of the Best Theory/Algorithm Paper Runner-Up Award at the IEEE ICDM07. He is a Fellow of the International Association of Pattern Recognition and the International Society of Optics and Photonics.



Chang Xu received the B.S. degree from TianJin University, Tianjin, China, in 2011. He is currently pursuing the Ph.D. degree with the Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China.

His current research interests include machine learning, information retrieval, and computer vision.

Chao Xu received the B.E. degree from Tsinghua University, Beijing, China, the M.S. degree from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 1988, 1991, and 1997, respectively.

He was an Assistant Researcher with the University of Science and Technology of China from 1991 and 1994. Since 1997, he has been with the Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, where he has been a Professor since 2005. He has authored or co-authored more than 100 papers in journals and conferences, and holds six patents. His current research interests include image and video processing, multimedia technology.

Hong Liu received the Ph.D. degree in mechanical electronics and automation from Peking University, Beijing, China, in 1996.

He is currently a Full Professor with the School of Electronics Engineering and Computer Science, Peking University, where he is the Director of the Engineering Laboratory on Intelligent Perception for Internet of Things. He has authored or co-authored more than 100 papers in journals and conferences. His current research interests include computer vision and robotics, image processing, and pattern recognition.

Dr. Liu was a recipient of the Chinese National Aero-space Award, the Excellence Teaching Award, and the Candidates of Top Ten Outstanding Professors from Peking University.

Yonggang Wen (S'99–M'08) received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, USA.

He is currently an Assistant Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. He has authored or co-authored more than 50 papers in journals and conferences. His system research on Cloud Social TV has been featured by international media, including The Straits Times, The Business Times, Lianhe Zaobao, Channel News Asia, ZDNet, CNet, United Press International, ACM Tech News, Times of India, Yahoo News. His current research interests include cloud computing, mobile computing, multimedia networks, cyber security, and green ICT.

Dr. Wen is a member of the Sigma Xi (the Scientific Research Society) and SIAM.