Research Article

Kernel-Based Multiview Joint Sparse Coding for Image Annotation

Miao Zang,^{1,2} Huimin Xu,¹ and Yongmei Zhang³

¹*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, No. 10, Xitucheng Road, Haidian District, Beijing 100876, China*

²*School of Electronics and Information Engineering, North China University of Technology, No. 5, Jinyuanzhang Road, Shijingshan District, Beijing 100144, China*

³*School of Computer Science, North China University of Technology, No. 5, Jinyuanzhang Road, Shijingshan District, Beijing 100144, China*

Correspondence should be addressed to Miao Zang; zangm@ncut.edu.cn

Received 28 October 2016; Accepted 11 January 2017; Published 19 March 2017

Academic Editor: Wanquan Liu

Copyright © 2017 Miao Zang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It remains a challenging task for automatic image annotation problem due to the semantic gap between visual features and semantic concepts. To reduce the gap, this paper puts forward a kernel-based multiview joint sparse coding (KMVJSC) framework for image annotation. In KMVJSC, different visual features as well as label information are considered as distinct views and are mapped to an implicit kernel space, in which the original nonlinear separable data become linearly separable. Then, all the views are integrated into a multiview joint sparse coding framework aiming to find a set of optimal sparse representations and discriminative dictionaries adaptively, which can effectively employ the complementary information of different views. An optimization algorithm is presented by extending K-singular value decomposition (KSVD) and accelerated proximal gradient (APG) algorithms to the kernel multiview framework. In addition, a label propagation scheme using the sparse reconstruction and weighted greedy label transfer algorithm is also proposed. Comparative experiments on three datasets have demonstrated the competitiveness of proposed approach compared with other related methods.

1. Introduction

With the surge of private and online images, automatic image annotation has become of great research interest in computer vision since it is a vital step for image retrieval and management [1, 2]. Image annotation aims to automatically predict a set of semantic labels, such as “sea,” “beach,” and “sand,” for the unannotated images by learning the relevance among images. However, the visual similar images may not be correlated with each other in semantics, which makes the problem still challenging.

In the past decade, considerable research efforts [3–12] have been made for automatic annotation. They can be roughly classified into three types of models: discriminative model, generative model, and nearest neighbor model. Discriminative model [3–5] treats image annotation as a

multilabel classification problem, in which each label is a single class. Given a test image, a semantic label will be propagated to it only if a corresponding classifier makes a decision that the image belongs to it. But this type of model neglects the correlation between different labels, which is also very important for image annotation. Generative model [6, 7] attempts to infer the correlations or joint probabilities between images and semantic concepts. By learning the statistical models, the test image is annotated based on probability computation. However, there exist many parameters in these models, which leads to heavy computational cost for parameter estimation. The nearest neighbor (NN) based model [8–12] predicts labels by taking a weighted combination of the label absence or presence among neighboring images. Because of its simplicity and efficiency, NN-based methods attract more researching attention.

Recently, sparse coding scheme and its variations have been successfully used in image annotation task. It is substantially related to the NN-based model since the test image is often represented by a few of representative samples in a low-dimensional manifold. For example, Wang et al. [13] reconstructed the test image sparsely and transferred labels by the sparse coefficients. Gao et al. [14] considered each (sub)class of images as a (sub)group and employed multilayer group sparsity to classify and annotate single-label images concurrently. Cao et al. [15] utilized the group sparse reconstruction framework to learn the label correlation for dictionary and reconstructed the test image for label prediction under weakly supervised case. Lu et al. [16] presented a more descriptive and robust visual bag-of-words (BOW) representation by semantic sparse recoding method for image annotation and classification. Jing et al. [17] learned a label embedded dictionary as well as a sparse representation for image annotation. In addition, the sparse representation can be further enhanced by exploiting the kernel mapping, since it maps nonlinear separable features into a higher dimensional feature space, in which features with the similar labels are closely grouped together while those without the same labels become linearly separable. Moran and Lavrenko [18] introduced a sparse kernel learning framework for the continuous relevance model, which can adaptively select different kernels for different features, and obtained great improvement.

Despite these efforts, most of existing annotation methods combine the information from different image features by concatenating them into a long feature vector, which treats all the features equally and omits their different contributions for the final decision. To solve this problem, Kalayeh et al. [19] introduced a multiview learning technique in image annotation, in which each type of features as well as label matrix is considered as a view, and all the views are adaptively integrated to exploit the complementary information, but the sparsity prior to the training samples used to reconstruct the test image is not considered. Liu et al. [20] introduced a multiview sparse coding framework for semisupervised image annotation. But they assumed the different views share a common sparse pattern, which ignores the diversity between the views. Yuan et al. [21] adopted a multitask joint sparse representation for image classification, allowing different coefficient representations for different task and enforcing the similarity among different tasks by joint sparsity. But image annotation is a multilabel classification problem, which cannot use the framework directly.

Motivated by the previous research, we formulate image annotation as a kernel-based multiview joint sparse coding (KMVJSC) learning problem. Figure 1 describes the framework of KMVJSC. Particularly, we map the feature views and label views to an implicit high-dimensional space and integrate all the views adaptively by multiview learning to strengthen the power of multiple modalities from an image. We aim to find a set of optimal sparse representations as well as dictionaries for each view simultaneously. Different views in an image should have similar coding coefficients to joint represent the same image on the one hand; on the other hand, these coefficients from different views should have some diversity to reflect the distinctive property of different views.

Thus, we adopt different sparse coefficients for different views, allowing each view to be flexibly coded over its associated dictionary; at the same time, we employ the joint sparsity constraint to make the sparse coefficients among different views to be similar. The optimization algorithm and label prediction scheme under the proposed framework are developed. Given a test image, we also map its multiple features into the same kernel space and reconstruct the test image by joint sparse coding using the learned atom representation dictionary. The product of the atom representation dictionary and the corresponding sparse coefficients is considered as scores of the near neighbors for the test image, and a greedy label transfer scheme is used to get the annotation. Experiments on three datasets demonstrate the effectiveness of our proposed method and the competitive performance compared with the related methods.

In summary, the major contributions of this paper are as follows: (1) An effective kernel-based multiview joint sparse coding framework is proposed and successfully applied in image annotation; (2) the optimization algorithm is proposed by extending the accelerated proximal gradient (APG) and K-singular value decomposition (KSVD) algorithms into a kernel-based multiview case; (3) a label prediction algorithm is proposed for kernel sparse representation framework based on the sparse reconstruction and weighted greedy label transfer scheme.

The rest of this paper is organized as follows: Section 2 briefly discusses the related work. Section 3 describes the details of our kernel-based multiview joint sparse representation, optimization algorithm, and label prediction scheme. Experiment results are reported and analyzed in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

In this section, we will review the related work including sparse representation, kernel sparse learning, and multiview learning.

2.1. Sparse Representation Based Image Annotation. Sparse coding aims to represent the observed data as a linear combination of dictionary entries (or training samples), with the constraint that each image feature vector is only represented by a subset of all the available dictionary atoms. Denote the feature matrix formed by original N training samples as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N] \in \mathbf{R}^{P \times N}$, where $\mathbf{X}_i = [x_{i1}, x_{i2}, \dots, x_{iP}]^T$ ($i = 1, 2, \dots, N$) is the feature vector of the i th sample image and P is the feature dimension. Then, the conventional sparse representation aims to seek a sparse coefficient matrix $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_N] \in \mathbf{R}^{N_d \times N}$ and the associated dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{N_d}] \in \mathbf{R}^{P \times N_d}$ by solving the following optimization problem [22]:

$$\arg \min_{\mathbf{D}, \mathbf{W}} \sum_{i=1}^N (\|\mathbf{X}_i - \mathbf{D}\mathbf{W}_i\|_F^2 + \lambda \|\mathbf{W}_i\|_1), \quad (1)$$

where $\|\mathbf{W}_i\|_1$ is the L_1 -norm sparse constraint regularization and λ is a trade-off parameter used to balance the sparsity and the reconstruction error.

FIGURE 1: Schematic illustration of the proposed KMVJSC approach for automatic image annotation. First, we extract multiple features from all the training images; each type of feature as well as the label matrix is treated as a view. Then, we map the different views into a kernel space, where joint sparse coding and dictionary learning are conducted. To label a test image, we extract its multiple features and reconstruct the test image using the learned atom representation dictionary in the mapped space by joint sparse coding. Finally, the product of atom representation dictionary and the corresponding sparse coefficients gives score for the near neighbors of the test image, and a greedy label transfer scheme is used to get the annotation.

For label transfer, given a test image, [13] adopts all the training images as dictionary and propagates the labels from the training images to the test image directly by the product of label matrix of training images with the learned sparse vector. Document [14] predicts the test image label based on the reconstruction error in the (sub)groups and assigns labels from the (sub)groups with the minimum reconstruction error to the test image. But the discriminative power of sparse representation in the original space is still limited compared to the kernel space.

2.2. Kernel Sparse Representation. The kernel sparse coding improves (1) by introducing a kernel trick on both training images and dictionary and has obtained great success in image classification. Let $\phi(\mathbf{X})$ denote the matrix whose columns are obtained by embedding the input features

$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N] \in \mathbf{R}^{P \times N}$ into a feature space using the mapping ϕ . That is, $\phi(\mathbf{X}) = [\phi(\mathbf{X}_1), \phi(\mathbf{X}_2), \dots, \phi(\mathbf{X}_N)]$. Furthermore, let $\phi(\mathbf{D}) = [\phi(\mathbf{D}_1), \phi(\mathbf{D}_2), \dots, \phi(\mathbf{D}_{N_d})]$ denote the dictionary in this mapped feature space; the kernel sparse representation [23] problem is defined as follows:

$$\arg \min_{\mathbf{D}, \mathbf{W}} \sum_{i=1}^N (\|\phi(\mathbf{X}_i) - \phi(\mathbf{D}) \mathbf{W}_i\|_F^2 + \lambda \|\mathbf{W}_i\|_1). \quad (2)$$

Equation (2) can be rewritten as in (3), which depends on a Mercer kernel function $\kappa(\cdot, \cdot)$, but not the mapping ϕ :

$$\arg \min_{\mathbf{D}, \mathbf{W}} \sum_{i=1}^N (\kappa(\mathbf{X}_i, \mathbf{X}_i) + \mathbf{W}_i^T \mathbf{K}_{\mathbf{D}\mathbf{D}} \mathbf{W}_i - 2 \mathbf{W}_i^T \mathbf{K}_{\mathbf{D}}(\mathbf{X}_i) + \lambda \|\mathbf{W}_i\|_1), \quad (3)$$

where $\mathbf{K}_{\mathbf{D}\mathbf{D}} \in \mathbf{R}^{N_d \times N_d}$ is a positive semidefinite matrix with the entries computed from Mercer kernel: $\{\mathbf{K}_{\mathbf{D}\mathbf{D}}\}_{ij} = \{\langle \phi(\mathbf{D}_i), \phi(\mathbf{D}_j) \rangle\}_{ij} = \kappa(\mathbf{D}_i, \mathbf{D}_j)$, and $\mathbf{K}_{\mathbf{D}}(\mathbf{X}_i)$ is a N_d -dimension vector with $\{\mathbf{K}_{\mathbf{D}}(\mathbf{X}_i)\}_j = \kappa(\mathbf{D}_j, \mathbf{X}_i)$. A Mercer kernel $\kappa(\mathbf{D}_i, \mathbf{D}_j)$ is a function representing the nonlinear similarity between two vectors $\mathbf{D}_i$ and $\mathbf{D}_j$. Some commonly used kernels include polynomial kernels, Gaussian kernels, and histogram intersection kernels (HIK).

For image classification, given a test image, it is assigned to the class with the minimum reconstruction error obtained by the learned dictionary and sparse coefficients for that class in the mapped space. However, image annotation task is a multilabel classification problem and cannot use this judgment decision for label transfer directly. In addition, the label transfer scheme [13] in original space cannot be moved to the mapped space since the mapping function is unknown. Thus, the application of kernel sparse representation in image annotation is limited.

2.3. Multiview Learning. While the previous framework has been proven successful for many tasks, it has only been applied to the single-view case. With multiple types of input modalities, [20] proposed a multiview learning based sparse representation model, which is based on the following general framework:

$$\arg \min_{\mathbf{D}^{(v)}, \mathbf{W}} \sum_{i=1}^N \left(\sum_{v=1}^V \|\mathbf{X}_i^{(v)} - \mathbf{D}^{(v)} \mathbf{W}_i\|_F^2 + \lambda \|\mathbf{W}_i\|_1 \right). \quad (4)$$

Here, each image sample is represented by V different features of views, $\mathbf{X}^{(v)}$ and $\mathbf{D}^{(v)}$ are the feature matrix for all the training images and learned dictionary of the v th view, respectively, and $\mathbf{W}$ is the sparse coefficient matrix shared for all views.

In [20], labels are treated as an additional $(V+1)$ th view; then, $\mathbf{X}^{(V+1)}$ and $\mathbf{D}^{(V+1)}$ represent the label view matrix of training samples and the associated label view dictionary, respectively. For a new test sample $\mathbf{y} = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(V)}\}$, the corresponding sparse code $\mathbf{w}$ can be obtained by solving the convex problem:

$$\arg \min_{\mathbf{w}} \sum_{v=1}^V \|\mathbf{y}^{(v)} - \mathbf{D}^{(v)} \mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 \quad (5)$$

and the label view of the test sample $\mathbf{y}^{(V+1)}$ is then predicted by $\mathbf{y}^{(V+1)} = \mathbf{D}^{(V+1)} \mathbf{w}$ directly. Although [20] implements the image annotation effectively, it adopts shared sparse coefficients for all views, which is not the case in reality because of the diversity of different views.

3. The Proposed Method

In this section, we introduce in detail our proposed method, optimization procedure, and label propagation algorithm. Throughout this paper, given a matrix $\mathbf{M}$, we will use the term $\mathbf{M}_i$ to denote its i th column vector and $\mathbf{M}_{i\cdot}$ to denote its i th row vector.

3.1. Kernel-Based Multiview Joint Sparse Representation. Motivated by the previous works of [20, 21], we utilize different sparse coefficient representation for different views, allowing the flexibility of sparse coefficients from distinct views; then, we introduce a joint sparsity constraint to our kernel-based multiview sparse learning framework to keep the correlation among multiple views. Specifically, let $\{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^V\}$ be a set of N training image samples obtained from V different feature views; each view contains the sample feature denoted by $\mathbf{X}^v = [\mathbf{X}_1^v, \mathbf{X}_2^v, \dots, \mathbf{X}_N^v] \in \mathbf{R}^{P_v \times N}$ ($v = 1, 2, \dots, V$), where $\mathbf{X}_i^v \in \mathbf{R}^{P_v}$ is the i th sample. The label information can be considered as another view $\mathbf{X}^{V+1} = [\mathbf{X}_1^{V+1}, \mathbf{X}_2^{V+1}, \dots, \mathbf{X}_N^{V+1}] \in \mathbf{R}^{P_{V+1} \times N}$, P_{V+1} is the number of labels, $\mathbf{X}_i^{V+1} \in \mathbf{R}^{P_{V+1}}$ is the label vector of the i th image, and each entry is either 1 or 0 representing the presence or absence of a given label in the image. Let $\mathbf{D}^v = [\mathbf{D}_1^v, \mathbf{D}_2^v, \dots, \mathbf{D}_{N_d}^v] \in \mathbf{R}^{P_v \times N_d}$ be an overcomplete dictionary ($N_d > P_v$), and $\mathbf{D}_j^v$ denotes the j th column of $\mathbf{D}^v$. Let $\phi(\mathbf{X}^v)$ and $\phi(\mathbf{D}^v)$ denote the corresponding sample matrix and dictionary of the v th view obtained using the mapping ϕ , respectively. Thus, we have $\phi(\mathbf{X}^v) = [\phi(\mathbf{X}_1^v), \phi(\mathbf{X}_2^v), \dots, \phi(\mathbf{X}_N^v)]$ and $\phi(\mathbf{D}^v) = [\phi(\mathbf{D}_1^v), \phi(\mathbf{D}_2^v), \dots, \phi(\mathbf{D}_{N_d}^v)]$. Then, the object function of kernel-based multiview joint sparse representation can be defined as

$$\arg \min_{\mathbf{D}^v, \Omega_i} \sum_{i=1}^N \left(\sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{D}^v) \mathbf{W}_i^v\|_F^2 + \lambda \|\Omega_i\|_{1,2} \right), \quad (6)$$

where $\mathbf{W}_i^v \in \mathbf{R}^{N_d}$ corresponds to the sparse representation coefficient of $\mathbf{X}_i^v$ over dictionary $\mathbf{D}^v$ in the mapped space, $\Omega_i = [\mathbf{W}_i^1, \mathbf{W}_i^2, \dots, \mathbf{W}_i^{V+1}] \in \mathbf{R}^{N_d \times (V+1)}$, $0 \leq i \leq N$, and $\|\Omega_i\|_{1,2}$ is defined as the sum of the L_1 -norm of all rows of the matrix Ω_i , which encourages Ω_i to be sparse in column direction and dense in row direction. It helps to enforce the sparse coefficients of different views to share a similar pattern. λ is the tuning parameter to control the regularization term.

The dictionaries from different views must be correlated with each other since they are learned from the same training images. To keep the correlation, we employ $\mathbf{D}^v = \mathbf{X}^v \mathbf{A}$, since dictionary atoms lie within the subspace spanned by the input data, where $\mathbf{A}$ is called the atom representation dictionary and $\mathbf{X}^v$ is called base dictionary [24]. Then, (6) can be rewritten as

$$\arg \min_{\mathbf{A}, \Omega_i} \sum_{i=1}^N \left(\sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}_i^v\|_F^2 + \lambda \|\Omega_i\|_{1,2} \right). \quad (7)$$

Compared with the dictionary obtained by using all the training samples directly, in which each dictionary atom corresponds to a training sample, the learned dictionary in (6) has no explicit physical meaning in their structure; that is, atoms located in the same column of different dictionary views may not originate from the same training sample, and also the correlations among different views are lost. By introducing $\phi(\mathbf{D}^v) = \phi(\mathbf{X}^v) \mathbf{A}$, we can learn a shared implicit dictionary $\mathbf{A}$ among multiple views in a similar way as (6), while the explicit nature (atom location information) of the dictionary is also maintained in the base dictionary $\phi(\mathbf{X}^v)$,

which is very important for subsequent label transfer in kernel space.

3.2. Optimization. To solve the problem in (7), we adopt an iterative optimization procedure by alternately optimizing technique [25] with respect to Ω_i ($i = 1, 2, \dots, N$) and $\mathbf{A}$ while holding the other one fixed. In the following, we provide a brief description of the alternating optimization for our method. For convenience, Notation lists the important notations used in this paper.

First, keeping $\mathbf{A}$ fixed, the problem in (7) can be simplified as

$$\arg \min_{\Omega_i} \sum_{i=1}^N \left(\sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}_i^v\|_F^2 + \lambda \|\Omega_i\|_{1,2} \right). \quad (8)$$

In the next step, keeping Ω_i fixed, the problem in (7) is simplified as

$$\arg \min_{\mathbf{A}} \sum_{i=1}^N \left(\sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}_i^v\|_F^2 \right), \quad (9)$$

which can also be represented by the matrix form:

$$\arg \min_{\mathbf{A}} \sum_{v=1}^{V+1} \|\phi(\mathbf{X}^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}^v\|_F^2. \quad (10)$$

Here, $\mathbf{W}^v = [\mathbf{W}_1^v, \mathbf{W}_2^v, \dots, \mathbf{W}_N^v] \in \mathbf{R}^{N_d \times N}$.

Such process can be iterated until the solutions of Ω_i and $\mathbf{A}$ converge to some local minimum. The following is a detailed description for optimization algorithm solving problems (8) and (10).

Learning Sparse Codes. Equation (8) can be decoupled into N distinct subproblems; the i th ($1 \leq i \leq N$) subproblem is formulated as follows:

$$\arg \min_{\Omega_i} \sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}_i^v\|_F^2 + \lambda \|\Omega_i\|_{1,2}. \quad (11)$$

The objective function in (11) is a nonsmooth convex function since $\|\Omega_i\|_{1,2}$ is nondifferentiable, which can be solved by accelerated proximal gradient (APG) method [26, 27] with the extension that signals are now in the multiview kernel space with a very high dimension.

Then, the optimization process can be iterated alternately by generalized gradient mapping and aggregation steps [27].

In the generalized gradient mapping step, we denote $\mathbf{K}_{\mathbf{X}\mathbf{X}}^v = \phi(\mathbf{X}^v)^T \phi(\mathbf{X}^v)$, $\mathbf{K}_{\mathbf{X}}^v(\mathbf{X}_i) = \phi(\mathbf{X}^v)^T \phi(\mathbf{X}_i^v)$, and $f(\Omega_i) = \sum_{v=1}^{V+1} \|\phi(\mathbf{X}_i^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{W}_i^v\|_F^2$; then, the gradient of $f(\Omega_i)$ with respect to Ω_i is calculated as follows:

$$\begin{aligned} \nabla f(\Omega_i) &= \left[\frac{\partial f}{\partial \mathbf{W}_i^1} \quad \frac{\partial f}{\partial \mathbf{W}_i^2} \quad \dots \quad \frac{\partial f}{\partial \mathbf{W}_i^V} \quad \frac{\partial f}{\partial \mathbf{W}_i^{V+1}} \right] \\ &= 2 \left[\mathbf{A}^T \mathbf{K}_{\mathbf{X}\mathbf{X}}^1 \mathbf{A} \mathbf{W}_i^1 - \mathbf{A}^T \mathbf{K}_{\mathbf{X}}^1(\mathbf{X}_i), \mathbf{A}^T \mathbf{K}_{\mathbf{X}\mathbf{X}}^2 \mathbf{A} \mathbf{W}_i^2 \right. \\ &\quad \left. - \mathbf{A}^T \mathbf{K}_{\mathbf{X}}^2(\mathbf{X}_i), \dots, \mathbf{A}^T \mathbf{K}_{\mathbf{X}\mathbf{X}}^{V+1} \mathbf{A} \mathbf{W}_i^{V+1} \right. \\ &\quad \left. - \mathbf{A}^T \mathbf{K}_{\mathbf{X}}^{V+1}(\mathbf{X}_i) \right]. \end{aligned} \quad (12)$$

Then, we apply the gradient mapping to (11) to get the following equivalent objective function:

$$\begin{aligned} &\arg \min_{\Omega_i} f(\Omega_i) + \lambda \|\Omega_i\|_{1,2} \\ &= \arg \min_{\Omega_i} f(\mathbf{U}) + \langle \nabla f(\mathbf{U}), \Omega_i - \mathbf{U} \rangle \\ &\quad + \frac{\eta}{2} \|\Omega_i - \mathbf{U}\|_F^2 + \lambda \|\Omega_i\|_{1,2}. \end{aligned} \quad (13)$$

It comprises the regularization term $\lambda \|\Omega_i\|_{1,2}$ and the approximation of $f(\Omega_i)$ by the first-order Taylor expansion at $\mathbf{U}$ regularized as the Euclidean distance between Ω_i and $\mathbf{U}$, where η is a parameter controlling the step penalty.

In the $t + 1$ th iteration, we calculate $(\Omega_i)_{t+1}$ as

$$\begin{aligned} (\Omega_i)_{t+1} &= \arg \min_{\Omega_i} f(\mathbf{U}_t) + \langle \nabla f(\mathbf{U}_t), \Omega_i - \mathbf{U}_t \rangle \\ &\quad + \frac{\eta}{2} \|\Omega_i - \mathbf{U}_t\|_F^2 + \lambda \|\Omega_i\|_{1,2}. \end{aligned} \quad (14)$$

Document [26] gives the solution for solving the optimization problem of (14) as

$$((\Omega_i)_{j,\cdot})_{t+1} = \left[1 - \frac{\lambda}{\eta \|\mathbf{Z}_{j,\cdot}\|_2} \right]_+ \mathbf{Z}_{j,\cdot}, \quad (15)$$

where $\mathbf{Z} = \mathbf{U}_t - (1/\eta) \nabla f(\mathbf{U}_t)$, $[\cdot]_+ = \max\{0, \cdot\}$, and $\mathbf{Z}_{j,\cdot}, (\Omega_i)_{j,\cdot}$ are the j th row of matrixes $\mathbf{Z}$ and Ω_i , respectively.

In the aggregation step, $\mathbf{U}_{t+1}$ is computed as a linear combination of $(\Omega_i)_{t+1}$ and $(\Omega_i)_t$:

$$\mathbf{U}_{t+1} = (\Omega_i)_{t+1} + \frac{1 - a_t}{a_t} a_{t+1} [(\Omega_i)_{t+1} - (\Omega_i)_t]. \quad (16)$$

Here, we set $a_{t+1} = 2/(t + 3)$ conventionally.

The steps of optimizing algorithm for (8) by the extended APG are elaborated as Algorithm 1.

Dictionary Updating. For problem (10), considering the matrices $\mathbf{X}_M = [(\mathbf{X}^1)^T, (\mathbf{X}^2)^T, \dots, (\mathbf{X}^{V+1})^T]^T \in \mathbf{R}^{\sum_{v=1}^{V+1} P_v \times N}$, $\mathbf{W}_M = [(\mathbf{W}^1)^T, (\mathbf{W}^2)^T, \dots, (\mathbf{W}^{V+1})^T]^T \in \mathbf{R}^{(V+1) \cdot N_d \times N}$, and $\mathbf{A}_M = [\mathbf{A}, \mathbf{A}, \dots, \mathbf{A}] \in \mathbf{R}^{N \times (V+1) \cdot N_d}$, we obtain

$$\mathbf{A}_M = \arg \min_{\mathbf{A}_M} \|\phi(\mathbf{X}_M) - \phi(\mathbf{X}_M) \mathbf{A}_M \mathbf{W}_M\|_F^2. \quad (17)$$

This is equivalent to the dictionary update stage in kernel dictionary learning algorithms. We adopt the kernel KSVD dictionary update strategy in [24] to efficiently solve it. Then, $\mathbf{A}$ is chosen as the first N_d column of $\mathbf{A}_M$.

3.3. Label Prediction. The proposed method treats labels as an additional view, so we can infer the label information from the sparse coefficients. In particular, given a test image represented by multiple feature views $\{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^V\}$, visual features of training images $\mathbf{X}^v$, the learned dictionary $\mathbf{A}$ from the features, and labels of the training images, we obtain the sparse coefficient vector $\{\omega^1, \omega^2, \dots, \omega^V\}$ ($\omega^V \in \mathbf{R}^{N_d}, v = 1, 2, \dots, V$) for the test image in terms of learned dictionary

Input:

Training image samples $\mathbf{X}^v$, ($v = 1, 2, \dots, V + 1$)
 Corresponding dictionary $\mathbf{A}$,
 Regularization parameter λ and step size η .

Output:

Sparse coefficient matrix $\mathbf{\Omega}_i$, ($i = 1, 2, \dots, N$)

(1) **For** $i = 1 : N$

(2) **Initialization:** Initialize $(\mathbf{\Omega}_i)_0 \in \mathbf{R}^{N_d \times (V+1)}$, $\mathbf{U}_0 = (\mathbf{\Omega}_i)_0$ to be zero matrix; $a_0 = 1$; and set $t = 0$.

(3) **While** not converged **Do**

Generalized gradient mapping step:

(4) Calculate $\mathbf{Z} = \mathbf{U}_t - (1/\eta)\nabla f(\mathbf{U}_t)$ by Eq. (12)

(5) **For** $j = 1 : N_d$

(6) Calculate j th row of $(\mathbf{\Omega}_i)_{t+1}$ by Eq. (15)

(7) **End For**

Aggregation step:

(8) $a_{t+1} = 2/(t + 3)$

(9) $\mathbf{U}_{t+1} = (\mathbf{\Omega}_i)_{t+1} + ((1 - a_t)/a_t)a_{t+1}[(\mathbf{\Omega}_i)_{t+1} - (\mathbf{\Omega}_i)_t]$

(10) $t = t + 1$

(11) **End While**

(12) **End For**

ALGORITHM 1: Optimizing method for problem (8).

Input:

The feature view of test image $\mathbf{y}^v \in \mathbf{R}^{P_v}$,
 Feature view of training images $\mathbf{X}^v$, ($v = 1, 2, \dots, V$),
 Learned dictionary $\mathbf{A} \in \mathbf{R}^{N \times N_d}$,
 Tuning parameter λ .

Output:

Predicted labels $\mathbf{y}^{V+1}$ for the test image.

(1) Calculate $\mathbf{\omega}^v$ ($v = 1, 2, \dots, V$) by Eq. (18)

(2) **For** $v = 1, 2, \dots, V$

(3) Find the five largest values in $\mathbf{A}\mathbf{\omega}^v$ and corresponding indices

(4) Find sample label column vectors in $\mathbf{X}^{V+1}$ corresponding to these indices

(5) **End For**

(6) Compute and rank the weighted frequency of labels appeared in all the found samples, with weights equaling to the product coefficients of $\mathbf{A}\mathbf{\omega}^v$

(7) Transfer labels according to their calculated frequency.

ALGORITHM 2: Label prediction procedure.

by solving the following convex problem, which can be solved similarly as to problem (11).

$$\begin{aligned} \arg\min_{\mathbf{\omega}^v} \sum_{v=1}^V \|\phi(\mathbf{y}^v) - \phi(\mathbf{X}^v) \mathbf{A} \mathbf{\omega}^v\|_F^2 \\ + \lambda \|\mathbf{\omega}^1 \ \mathbf{\omega}^2 \ \dots \ \mathbf{\omega}^V\|_{1,2}. \end{aligned} \quad (18)$$

Since the magnitude of the product term $\mathbf{A}\mathbf{\omega}^v$ can be considered as the importance of certain training samples in the reconstruction of the test image, the priority of the samples used for label propagation is based on the magnitude of its corresponding product term $\mathbf{A}\mathbf{\omega}^v$. Although $\mathbf{\omega}^{V+1}$ is unknown, it must share the similar pattern with $\mathbf{\omega}^v$ ($1 \leq$

$v \leq V$), since they are used to represent the same image. So, we iterate for each view v , choosing the samples in $\mathbf{X}^{V+1}$ corresponding the top five values in $\mathbf{A}\mathbf{\omega}^v$. Then, the tag propagation for the test image is based on a weighted version of greedy label transfer scheme [8]. The proposed label prediction and propagation scheme is summarized in Algorithm 2.

4. Experiments

To validate the effectiveness of the proposed KMOVJSC for automatic image annotation task, we conduct experiments and compare the results with related algorithms on three popular benchmark databases.

4.1. Experimental Settings. The three datasets, we used to evaluate the performance of our approach are as follows.

Corel5K dataset is the most popular dataset for annotation evaluation [8–10, 13, 17–19, 28, 29]. It consists of 5000 images from 50 different topics, such as “beach,” “aircraft,” and “horse,” and each topic includes 100 similar images. The training and testing sets contain 4500 and 500 images, respectively. Each image is manually annotated with 1 to 5 labels from a dictionary of 260 labels, and with 3.5 labels on average.

IAPRTC12 [30] dataset consists 19627 images of natural scenes dealing with sports and actions and photographs of people, animals, cities, landscapes, and so on. In the dataset, 17665 images are selected for training and the remaining 1962 images are chosen for testing. The number of labels in this dataset is 291, and each image has up to 23 labels with an average of 5.7 labels per image. Besides, each label averagely relates with 347 images.

MIRFlickr25K [31] contains 25000 images which are collected from the social photography site <https://www.flickr.com> and equally split for training and test. It provides 38 labels as the ground-truth annotation, such as “animals,” “baby,” and “baby*.” The words with “*” indicate a further strict annotation for labels. Each image is associated with up to 17 labels and 4.7 labels on average, while each label averagely relates with 1560 images. Besides, the dataset provides 1386 tags. Since the Flickr tags are noisy, we kept the tags that appear at least 50 times, resulting in a vocabulary of 457 tags, and use it as another type of feature.

We adopt histogram intersection kernel as the kernel function because it is a parameter-free kernel, and it has obtained excellent performance in evaluating the similarity between two histograms [14].

Following [32], we adopt 15 visual features but limit the size of large feature vectors. Specifically, in order to reduce the computation complexity of both training and testing procedures, we use color histogram with 10 bins per color channel. For features encoded with the spatial layout information, we quantize the color with 8 bins in each channel and reduce the k -means cluster centers to 500 for SIFT features. Thus, the 15 distinct features include 3 color features (RGB, LAB, HSV, 1000D), 2 SIFT features (DenseSIFT, HarrisSIFT, 1000D), 2 Hue features (DenseHue, HarrisHue, 100D), 7 above features with layout encoding (RGBV3H1, LABV3H1, HSVV3H1, 1536D; DenseSIFTV3H1, HarrisSIFTV3H1, 1500D; DenseHueV3H1, HarrisHueV3H1, 300D), and a GIST feature (512D). In addition, for MIRFlickr25K dataset, we endow each image with a binary vector indicating the absence or presence of each tag, which leads to a Tag feature of length 457D.

There are two parameters that need to be tuned, that is, dictionary size N_d and regularization parameter λ . In this paper, N_d is selected from {2000, 2500, 3000, 3500, 4000, 4500}, and λ is selected from {0.0001, 0.001, 0.01, 0.1, 1, 10}. We perform KMVJSC with fivefold cross validation on training set and conduct sensitivity analysis in Section 4.3. Finally, we set the parameters of our KMVJSC algorithm as $N_d = 2000$, $\lambda = 0.1$ on Corel5K dataset and as $N_d = 3000$, $\lambda = 0.001$ on the other two datasets. Due to the

random entries in initialization, we repeat all the experiments 5 times separately and report the average results.

Following most research [10, 13, 17, 19, 32], all the test images are annotated by the top 5 relevant labels. To estimate the performance, we calculate precision (P), recall (R), and F_1 -measure for each label. For a given label c , the three metrics are defined by $R = N_c/N_g$, $P = N_c/N_t$, and $F_1 = 2 \times R \times P/(R + P)$. Here, N_g is the number of images labeled with c in the ground truth, N_t is the number of images labeled with c by our automatic annotation algorithm, and N_c is the number of correct labeled images with c . We present the mean value over all labels for each metric. Besides, the number of labels with nonzero recall (N_+) is also used.

4.2. The Compared Methods. We compare our KMVJSC with some sparse coding based baselines including multilabel sparse coding (MSC) [13], kernel sparse coding (KSC) [23], multiview Hessian discriminative sparse coding (MVHDS) [20], multiview joint sparse coding (MVJSC) in original space, and kernel-based multiview sparse coding (KMVSC) without enforcing the joint sparsity across views. For single-view methods (MSC and KSC), we concatenate the multiple features as a long vector. For KSC, we use the same label transfer scheme as our method except using learned dictionary and reconstruct coefficient directly; for MVJSC, we calculate the average sparse coefficient from multiple feature views and use it as the sparse coefficient of label view and transfer labels with the same way as that of MVHDS; and for KMVSC, we use the same label transfer method as our KMVJSC. Moreover, we take the same parameter selection strategy for these methods, where all the training, validation, and test sets are exactly the same as those used for our method.

4.3. Experimental Results. Table 1 lists some image examples from three datasets and the predicted labels using our method, together with the ground truths. It contains at least one mismatched label compared with ground-truth labels (perfectly matched annotations are not listed here). The differences in predicted labels are marked in italic font. The results in Table 1 show that, in many cases, some predicted labels missed by the ground-truth annotation can still explain the image well, such as “meadow” for the first image from Corel5K dataset. Besides, some semantically similar words were also treated as errors, such as “cathedral” for the last image from IAPRTC12 dataset, which shows the potential effectiveness of our proposed method for automatic image annotation task.

The annotation performance of our proposed algorithms on three datasets along with other related methods is listed in Table 2, where the methods with “*” represent the implementation using our features, and KMVJSC+T refers to the annotation with our method using another Tag feature for MIRFlickr25K besides the 15 feature views and the label view. The maximum values are bold. From the result in Table 2, we can make the following observations.

(1) For single-view methods, KSC is slightly better than MSC; for multiview methods, kernel methods including KMVSC and KMVJSC both outperform MVJSC, which demonstrates the power of kernel mapping. We also can see

TABLE 1: Comparison of predicted labels with ground-truth annotations for images from Corel5K, IAPRTC12, and MIRFlickr25K datasets. Italic labels are the difference between predicted and ground-truth labels while some of them are also relevant to the image.

Images from Corel5K
Ground truths	Cow, bulls, antlers, elk	Grass, ruins, stone, column	Coral, anemone, ocean, reefs	Grass, bear, polar, tundra	Cars, tracks, turn, prototype
Predicted labels	<i>Meadow</i> , cow, bulls, elk, antlers	Ruins, stone, <i>palace</i> , column, grass	Coral, anemone, ocean, reefs, <i>flowers</i>	Grass, bear, polar, tundra, <i>tree</i>	Cars, tracks, turn, prototype, <i>birds</i>
Images from IAPRTC12
Ground truths	cloth, front, girl, hat, wall	court, man, player, tennis	beach, boat, fog, man, sea, tower	bush, cloud, group, hill, island, lake, people, sky	church, lamp, night, sky, square, street, tower
Predicted labels	front, girl, hat, <i>jacket</i> , wall	court, man, player, tennis, <i>net</i>	beach, boat, fog, man, <i>people</i>	group, hill, island, lake, <i>photo</i>	<i>cathedral</i> , lamp, night, square, street
Images from MIRFlickr25K
Ground truths	animals, bird, bird* , clouds, plant_life, sky, structures, sunset, tree	plant_life, structures, tree, tree*	flower, flower* , night, plant_life	animals, plant_life, structures, tree, water	car, car* , plant_life, sky, structures, transport, tree
Predicted labels	clouds, plant_life, sky, structures, tree	plant_life, sky, structures, tree, tree*	flower, flower* , plant_life, <i>indoor</i> , night	animals, plant_life, tree, sky, structures	indoor, structures, plant_life, sky, tree

TABLE 2: Performance comparison using different methods on Corel5K, IAPRTC12, and MIRFlick25K datasets. Methods with “*” means the implementation using our features. The best scores for each method are highlighted in bold. Using the same features, KMOVJSC get the best results on all the three datasets. KMOVJSC+T is the KMOVJSC method using an additional Tag feature on MIRFlick25K dataset.

Datasets	Corel5K				IAPRTC12				MIRFlick25K			
Method	AP	AR	F_1	N+	AP	AR	F_1	N+	AP	AR	F_1	N+
MSC* [13]	0.26	0.32	0.287	138	0.35	0.28	0.311	252	0.32	0.24	0.274	36
MVHDSC* [20]	0.35	0.37	0.360	160	0.45	0.32	0.374	260	0.40	0.30	0.343	38
KSC* [23]	0.29	0.33	0.309	143	0.36	0.30	0.327	256	0.33	0.24	0.278	37
KMVSC	0.35	0.38	0.364	154	0.44	0.33	0.377	262	0.39	0.29	0.333	38
MVJSC	0.34	0.36	0.350	150	0.42	0.32	0.363	261	0.37	0.29	0.325	38
KMOVJSC	0.37	0.4	0.384	162	0.46	0.34	0.391	268	0.41	0.32	0.359	38
KMOVJSC+T	—	—	—	—	—	—	—	—	0.43	0.33	0.373	38

that MVHDSC is still comparable with KMOVSC, which is probably because MVHDSC enforces structured sparsity on dictionary additionally.

(2) All multiview methods dramatically outperform the single-view methods on all datasets. Although MSC takes multilabel information into consideration, its performances are inferior to those of multiview methods including MVJSC, which shows multiview learning can harness the label view and feature views in a more natural way and capture the relation between multiple labels and that between labels and visual features for discrimination.

(3) Compared to KMOVSC, which neglects the joint sparsity across multiple views, our KMOVJSC clearly outperforms KMOVSC, which shows the constraint of joint sparsity across different views is valuable to enforce the robustness in coefficient estimation among different views.

(4) Adding the Flickr tags as features helps to improve the annotation performance further, which illustrates that tags are another type of complementary information to visual features and label view and can be effectively used in multiview learning.

4.4. Sensitivity Analysis of Parameters. We investigate the sensitivity of the parameters N_d and λ in our approach using the three datasets. The corresponding results are presented in Figures 2 and 3.

It should be noted that the results in these two figures are obtained by 5-fold cross validation on the training image set of the three datasets, respectively. Figure 2 illustrates the values of the F_1 measure versus different size of dictionary with fixing λ (0.1 for Corel5K dataset and 0.001 for the other two datasets). From the results, we can observe that although the performance in terms of F_1 measure varies on changing the value of N_d , it is not very sensitive to its choice and remains fairly stable. Considering that high dimension corresponds to expensive computing cost, we set $N_d = 2,000$ for Corel5K dataset and $N_d = 3,000$ for IAPRTC12 and MIRFlickr25K datasets to leverage the performance and the cost. Figure 3 shows the values of the F_1 measure versus different λ with the selected fixing N_d . It is observed that the $L_{1,2}$ -norm joint sparsity regularization term is effective while it is not too small or too large. If it is too small, the correlations between different views are lost while it limits the flexibility of all the variety of views with large values. We can see that

FIGURE 2: F_1 measure of our KMOVJSC with N_d changing on the training set of the three datasets. The horizontal axis represents the size of the dictionary, from 2000 to 4500 with a step size of 500. The vertical axis represents the performance of F_1 .

the best result of our approach is reached when $\lambda = 0.1$ on Corel5K dataset and $\lambda = 0.001$ on the other two datasets.

5. Conclusions

In this paper, we present a kernel-based multiview joint sparse coding framework for image annotation problems. It can learn a set of optimized sparse representations as well as dictionaries in a multiview kernel space. We consider label view (or tag view) as an additional view and adaptively utilize the relationship between label views and visual feature views to find the more discriminative sparse representation by incorporating multiview learning and the $L_{1,2}$ -norm joint sparsity. We extend KSVD and AGM to a multiview kernel version to solve our optimization problem. The priority of sparse coefficients in the kernel space is used to propagate labels to the test images by weighted greedy label transfer scheme. Experimental results on three widely used datasets

FIGURE 3: F_1 measure of our KMVJSC with λ changing on the training set of the three datasets. The horizontal axis represents the tuning parameter λ , from $1e-4$ to 10 with a step size of 10 times. The vertical axis represents the performance of F_1 .

show the effectiveness of our proposed method compared to the related methods for image annotation tasks. Considering that some non-ground-truth annotations can still describe the image well in our experiments, for the future research, we try to further involve human perception [33] to measure the relevancy between the non-ground-truth annotations and the images to show the potential flexibility of our proposed framework. Besides, we are going to apply deep learning features [10, 34] and multiple kernel learning [18] into our framework.

Notation

- N : Number of training samples with labels
- $\mathbf{X}^v$: Training samples data of the v th view
- $\mathbf{X}_i^v$: i th training sample of the v th view
- P_v : Dimension of v th view
- ϕ : Kernel mapping
- $\mathbf{D}^v$: Dictionary of the v th view
- V : Number of feature views
- N_d : Number of dictionary atoms
- $\mathbf{A}$: Atom representation dictionary
- $\mathbf{W}_i^v$: Sparse coefficient of $\mathbf{X}_i^v$ over $\mathbf{D}^v$ in ϕ space
- $\mathbf{W}^v$: $[\mathbf{W}_1^v, \mathbf{W}_2^v, \dots, \mathbf{W}_N^v]$
- $\mathbf{\Omega}_i$: $[\mathbf{W}_i^1, \mathbf{W}_i^2, \dots, \mathbf{W}_i^{V+1}]$
- λ : Parameter of regularization term
- $\mathbf{y}^v$: Test image data of the v th view
- $\mathbf{\omega}^v$: Sparse coefficient of $\mathbf{y}^v$ over $\mathbf{D}^v$ in ϕ space.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by Chinese National Natural Science Foundation (61371143), Beijing Natural Science Foundation (4132026), and Scientific Research Project of Beijing Municipal Education Commission (KM201510009006).

References

- [1] D. Zhang, M. M. Islam, and G. Lu, "A review on automatic image annotation techniques," *Pattern Recognition*, vol. 45, no. 1, pp. 346–362, 2012.
- [2] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. M. Snoek, and A. Del Bimbo, "Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys*, vol. 49, no. 1, article 14, 2016.
- [3] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
- [4] Y. Verma and C. V. Jawahar, "Exploring SVM for image annotation in presence of confusing labels," in *Proceedings of the 24th British Machine Vision Conference*, pp. 1–11, 2013.
- [5] M.-L. Zhang and L. Wu, "LIFT: multi-label learning with label-specific features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 107–120, 2015.
- [6] O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proceedings of the 9th International Workshop on Multimedia Data Mining*, pp. 1–7, Las Vegas, Nev, USA, August 2008.
- [7] D. Putthividhya, H. T. Attias, and S. S. Nagarajan, "Topic regression multi-modal Latent Dirichlet Allocation for image annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3408–3415, IEEE, San Francisco, Calif, USA, June 2010.
- [8] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proceedings of the 10th European Conference on Computer Vision*, pp. 316–329, Marseille, France, October 2008.
- [9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 309–316, IEEE, Kyoto, Japan, September–October 2009.
- [10] Y. Verma and C. V. Jawahar, "Image annotation by propagating labels from semantic neighbourhoods," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 126–148, 2017.
- [11] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1310–1322, 2009.
- [12] Z. Chen, J. Cao, T. Xia, Y. Song, Y. Zhang, and J. Li, "Web video retagging," *Multimedia Tools and Applications*, vol. 55, no. 1, pp. 53–82, 2011.
- [13] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proceedings of the IEEE 22nd International Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1643–1650, IEEE, Miami, Fla, USA, June 2009.
- [14] S. Gao, L.-T. Chia, I. W.-H. Tsang, and Z. Ren, "Concurrent single-label image classification and annotation via efficient multi-layer group sparse coding," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 762–771, 2014.

- [15] X. Cao, H. Zhang, X. Guo, S. Liu, and D. Meng, "SLED: semantic label embedding dictionary representation for multilabel image annotation," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2746–2759, 2015.
- [16] Z. Lu, P. Han, L. Wang, and J.-R. Wen, "Semantic sparse recoding of visual content for image applications," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 176–188, 2015.
- [17] X.-Y. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang, "Multi-label dictionary learning for image annotation," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2712–2725, 2016.
- [18] S. Moran and V. Lavrenko, "Sparse kernel learning for image annotation," in *Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR '14)*, pp. 113–120, ACM, Glasgow, UK, April 2014.
- [19] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: image annotation using weighted multi-view non-negative matrix factorization," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 184–191, IEEE, Columbus, Ohio, USA, June 2014.
- [20] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview Hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.
- [21] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.
- [22] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE 22nd International Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, IEEE, Miami, Fla, USA, June 2009.
- [23] S. Gao, I. W. Tsang, and L.-T. Chia, "Sparse representation with kernels," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 423–434, 2013.
- [24] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 2021–2024, IEEE, Kyoto, Japan, March 2012.
- [25] J. C. Bezdek and R. J. Hathaway, "Convergence of alternating optimization," *Neural, Parallel and Scientific Computations*, vol. 11, no. 4, pp. 351–368, 2003.
- [26] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proceedings of the IEEE 5th International Conference on Data Mining*, pp. 746–751, Las Vegas, Nev, USA, July 2009.
- [27] X.-T. Yuan and S. C. Yan, "Visual classification with multitask joint sparse representation," in *Proceedings of the IEEE 23rd International Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3493–3500, IEEE, San Francisco, Calif, USA, June 2010.
- [28] P. Duygulu, K. Barnard, J. F. G. Freitas, and D. A. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary," in *Proceedings of the 7th European Conference on Computer Vision*, pp. 97–112, Copenhagen, Denmark, May 2002.
- [29] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 42, no. 3, pp. 838–849, 2012.
- [30] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: a new evaluation resource for visual information systems," in *Proceedings of the International Conference on Language Resources and Evaluation*, Genoa, Italy, May 2006.
- [31] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, Vancouver, Canada, October 2008.
- [32] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the MIRFlickr set," in *Proceedings of the 11th ACM International Conference on Multimedia Information Retrieval*, pp. 537–546, Philadelphia, Pa, USA, March 2010.
- [33] V. Bruni and D. Vitulano, "An improvement of kernel-based object tracking based on human perception," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 11, pp. 1474–1485, 2014.
- [34] F. Wu, Z. Wang, Z. Zhang et al., "Weakly semi-supervised deep learning for multi-label image annotation," *IEEE Transactions on Big Data*, vol. 1, no. 3, pp. 109–122, 2015.