A Locality Sensitive Low-Rank Model for Image Tag Completion

Xue Li, Bin Shen, *Member, IEEE*, Bao-Di Liu, and Yu-Jin Zhang, *Senior Member, IEEE*

***Abstract*—Many visual applications have benefited from the outburst of web images, yet the imprecise and incomplete tags arbitrarily provided by users, as the thorn of the rose, may hamper the performance of retrieval or indexing systems relying on such data. In this paper, we propose a novel locality sensitive low-rank model for image tag completion, which approximates the global nonlinear model with a collection of local linear models. To effectively infuse the idea of locality sensitivity, a simple and effective pre-processing module is designed to learn suitable representation for data partition, and a global consensus regularizer is introduced to mitigate the risk of overfitting. Meanwhile, low-rank matrix factorization is employed as local models, where the local geometry structures are preserved for the low-dimensional representation of both tags and samples. Extensive empirical evaluations conducted on three datasets demonstrate the effectiveness and efficiency of the proposed method, where our method outperforms pervious ones by a large margin.**

***Index Terms*—Automatic image annotation, image tag completion, locality sensitive model, low-rank matrix factorization.**

I. Introduction

THE advent of the big data era has witnessed an explosive growth of the visual data, which has spawned many visual applications to organize, analyze, and retrieve these images. However, user-labeled visual data, such as images which are uploaded and shared in Flickr, are usually associated with imprecise and incomplete tags. This will pose threats to the retrieval or indexing of these images, causing them difficult to be accessed by users. Unfortunately, missing label is inevitable in the manual labeling phase, since it is infeasible for users to label every related word and avoid all possible confusions, due to the existence of synonyms and user preference. Therefore, image tag completion or refinement has emerged as a hot issue in the multimedia community.

Manuscript received March 21, 2015; revised August 9, 2015 and November 23, 2015; accepted January 2, 2016. Date of publication January 18, 2016; date of current version February 18, 2016. This work was supported by the National Nature Science Foundation under Grant NNSF: 61171118, and by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant SRFDP-20110002110057. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lexing Xie.

X. Li and Y.-J. Zhang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: xue-li11@mails.tsinghua.edu.cn; zhang-yj@tsinghua.edu.cn).

B. Shen was with Purdue University, West Lafayette, IN 47907 USA. He is now with Google Research, New York, NY 10011 USA (e-mail: stanshenbin@gmail.com).

B.-D. Liu is with the Department of Information and Control Engineering, China University of Petroleum, Qingdao 266580, China (e-mail: thu.liubaodi@gmail.com).

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

Digital Object Identifier 10.1109/TMM.2016.2518478

In the scenario of image tag completion, all the images are assumed to be partially labeled, for instance an image whose true labels are $\{c_1, c_2, c_3\}$ may only be labeled as $\{c_2\}$, while c_1 and c_3 are missing. The goal of image tag completion is to accurately recover the missing labels for all the images. A plethora of algorithms have been developed to address this issue, among which many researchers explore the insight that related tags are often concurrent with each other, and images depicting similar contents tend to have related tags. However, existing completion methods are usually founded on linear assumptions, hence the obtained models are limited due to their incapability to capture complex correlation patterns.

To enable nonlinearity and keep the computational efficiency at the same time, we resort to a locality sensitive approach, with the assumption that albeit nonlinear globally, the model can be linear locally, which allows the application of linear models when samples are restricted to individual regions of the data space. Following this idea, the entire data space is divided into multiple regions, within each of which a local linear model is learnt, leading to a model denoted as Locality Sensitive Low-rank Reconstruction (LSLR).

The first issue involving in such a locality sensitive framework is how to conduct meaningful data partition, which is nontrivial in the tag completion scenario, since the distance between samples, which is essential to most partition methods, is extremely unreliable when measured by low-level features and incomplete user-provided tags. To handle such issues, a simple and effective pre-processing module is designed, by eliminating the side effect of both high-frequency and rare tags, and learning for each sample the low-dimensional representation suitable for partition.

The second problem concerns the construction of the local models, that is, how to effectively model the local correlations between similar samples and related tags. In this paper, our method draws inspiration from Multi-Task Learning (MTL) and formulates the local models by low-rank matrix factorization [1], [2]. Specifically, each initial tag sub-matrix is decomposed into a low-rank basis matrix and a sparse coefficient matrix, and the compressed representation for both the tags and samples are learnt, respectively. Such a model is able to promote information sharing between related tags as well as similar images.

However, it is not preferable to learn local models independently, since the output of data partition is typically far from satisfactory, even with the help of the pre-processing module. As a result, the local models learned independently tend to overfit the data restricted to individual regions. Therefore, to relieve the risk of overfitting as well as to promote robustness of the

proposed LSLR method, a global consensus model is introduced to regularize the local models.

To our knowledge, we are the first to infuse the idea of locality sensitivity into the scenario of image tag completion, and our main contributions are summarized as follows.

1) We propose a locality sensitive low-rank model for image tag completion, which approximates the global nonlinear model with a collection of local linear models, by which complex correlation structures can be captured.
2) Several adaptations are introduced to enable the fusion of locality sensitivity and low-rank factorization, including a simple and effective pre-processing module and a global consensus regularizer to mitigate the risk of overfitting.

The remainder of this paper is organized as follows. Section II gives a brief overview of related studies. The proposed tag completion formulation is elaborated in Section III, followed by detailed optimization algorithms in Section IV. Evaluation results on three datasets are presented in Section V, and Section VI concludes this paper.

II. Related Work

Image tag completion, which aims at recovering missing tags of images, is actually a special case of automatic image annotation (AIA). To draw a parallel between the two topics, recent efforts on both areas are briefly reviewed below.

Given an unlabeled image, the goal of image annotation is to identify its contents and label it with an appropriate number of tags. Numerous methods have been proposed in this area, including mixture models such as MBRM [3], SML [4], topic models such as mmLDA, cLDA [5], tr-mmLDA [6], discriminative methods [7], and label-transfer schemes [8]. Among them, state-of-the-art performance is reported by label-transfer methods. Specifically, JEC [8] adopted equal weights for each feature and transferred labels in a greedy manner. TagProp [9] embedded metric learning to learn more discriminative weights. 2PKNN [10] extended LMNN [11] into a multi-label scenario and constructed semantic groups to boost annotation performance for rare tags.

Despite their success, a properly labeled training set is usually required for the above methods, which is unrealistic for large scale real world datasets. Therefore, several recent studies are conducted on developing annotation algorithms robust to missing labels, including [12]–[17]. Specifically, [12] proposed a ranking based multi-label learning framework, and handled missing labels by combining the ranking losses via a group lasso regularizer. Learning image annotation models from partially labeled training data is much more challenging than solving traditional AIA tasks, since the lack of fully labeled training set limits the leverage of some sophisticated supervised models, thus the annotation accuracy is far from satisfactory.

As an alternative, many researchers proposed to directly recover the missing labels via exploiting information from the incomplete initial tags. Significant efforts have been devoted to the task of image tag completion, among which many different approaches [18]–[23] have been explored from divergent perspectives. Specifically, [24] proposed a tag recommendation method comprising the generation and aggregation of candidate tags. In [25], tag recommendation was approached as a maximum a posteriori (MAP) problem using a folksonomy. G. Zhu *et al.* [26] decomposed the user-provided tag matrix into a low-rank completed matrix and a sparse error matrix. Similarly in [27], tag completion was handled via nonnegative matrix factorization. Alternatively, the TMC method [28] directly searched for the optimal tag matrix which preserved correlation structures for both images and tags. The recently proposed LSR method [29] conducted linear sparse reconstruction for each image and each tag, respectively. Though impressive progresses have been achieved, most of the aforementioned methods failed to consider the complex structures beyond the capability of linear models.

Methodologically, the idea of approximating a nonlinear model using a collection of local linear models has been explored in other areas as well. For instance, [30] proposed a novel locally linear SVM classifier, and recently in [31], a local collaborative ranking method was developed for recommendation systems. In this paper, to apply this strategy to image tag completion, several key components are introduced, as will be explained in detail in Section III.

III. Proposed LSLR Model

This section describes the proposed LSLR model for tag completion, as shown in Fig. 1.

A. Overview of the Locality Sensitive Framework

Assume we are given n partially labeled images, whose visual feature matrix and initial tag matrix is denoted as $X \in \mathbb{R}^{n\times d}$ and $D \in \mathbb{R}^{n\times m}$, respectively, where d is the dimension of visual feature, and m is the size of our vocabulary. Our goal for tag completion is to recover the complete tag matrix Y. The proposed method achieves this via several modules, including pre-processing, data partition, and the learning of local models. As sketched in Fig. 1(a), the low-dimensional representation is learnt for each sample in the phase of pre-processing. Based on this novel representation, all the images in the dataset are divided into multiple groups, so that samples within the same group are semantically related.

As illustrated in Fig. 1(b), a local model is then established by factorizing the complete matrix Y_i into a basis matrix W_i and a sparse coefficient matrix H_i, as shown below

$$Y_i = W_i H_i \quad \forall i \in 1, 2, \cdots, c, \tag{1}$$

where $W_i \in \mathbb{R}^{n_i\times k}$, $H_i \in \mathbb{R}^{k\times m}$, and n_i is the number of samples in the i-th cluster. Since $k \ll m$, the related tags are encouraged to be reconstructed by common basis in W_i, and thus tag-level information gets shared. Then our final completed matrix Y can be obtained by integrating all the sub-matrices Y_is.

Following the definition of (1), our locality sensitive low-rank model can be formulated as

$$f = \sum_{i=1}^{c} (L_i + \lambda R g_i) \tag{2}$$

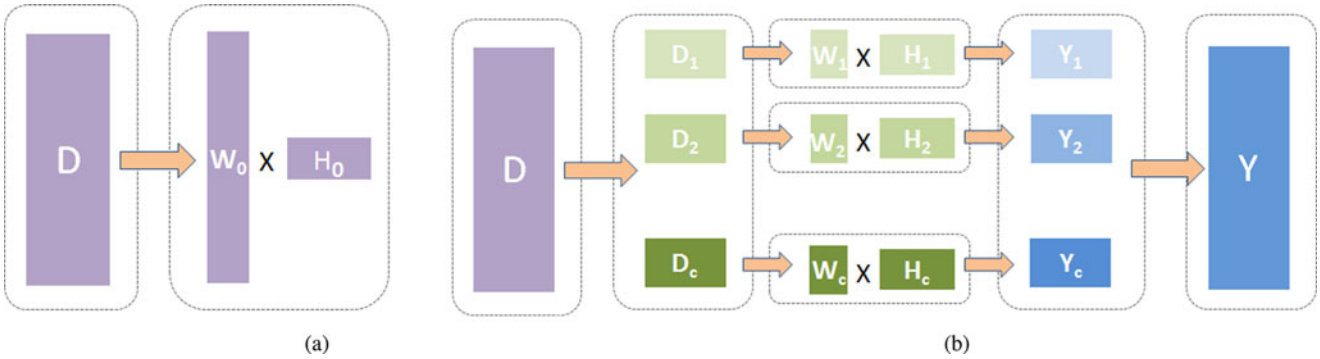

Fig. 1. Framework of the proposed LSLR. (a) shows our pre-processing module, which learns a low-dimensional image level representation (W_0) suitable for partition. (b) illustrates the locality sensitive framework, where the initial tag matrix D is partitioned into c clusters, then a local linear model is learnt for each cluster, through matrix factorization. The final completed matrix is obtained by integrating the resulted Y_is.

where L_i is the local model for the i-th cluster, and Rg_i denotes the global consensus regularizer imposed on the i-th cluster, with a trade-off parameter λ.

The loss function L_i can be further broken down into the following terms:

$$L_i = \|D_i - W_i H_i\|_F^2 + \eta R_{W_i} + \gamma R_{H_i} + 2\beta \|H_i\|_1 \quad (3)$$

where D_i is the initial tag matrix for cluster i, R_{W_i} and R_{H_i} are regularization for W_i and H_i, respectively, and η, γ and β are parameters. Concrete definitions of the items are presented in the following subsections.

B. Pre-Processing and Data Partition

This section introduces two closely related modules: pre-processing and data partition. As mentioned in Section III-A, the goal of data partition is to divide the entire sample space into a collection of local neighborhoods or groups, such that samples within each group are semantically related. However, as we observed in our experiments, direct partitions usually fail to generate meaningful groups, regardless of using visual features or incomplete initial tags. The reason behind is easy to understand. For instance, images depicting *people* may be divided into the clusters concerning *beach* or *building* according to their backgrounds, especially when *people* is missing. On the other hand, despite actually describing different contents such as *bear*, *fox* or *mountain*, samples initially labeled as *snow* may be grouped into the same cluster about *snow*, since distance is distorted when their foreground tags are absent.

In this paper, a cluster is referred to as a *cluttered cluster* if its images are not actually semantically related, and a *compact cluster* otherwise. To alleviate the risk of generating cluttered clusters, a two-step pre-processing module is employed to learn the low-dimensional representation that is less correlated, as shown in Fig. 1(a).

Our first step is to eliminate the side effect of both the high-frequency and rare tags by removing their corresponding columns in the initial tag matrix, since they hardly appear as the main content of the images. For instance, *sky* usually relates to background rather than foreground, but the learning process may consider it as an intrinsic pattern due to its high-frequency, thereby preserving its information in the low-dimensional representation. To identify tags that need to be removed, some thresholds are manually set based on the counts of the initial tags.[1]

The second step is to learn the low-dimensional representation for each image. Recall that the basis matrix in (1) can be interpreted as row-wise low-dimensional representation for each sample, thus it can be readily adapted to fit our demand. Specifically, we solve (3) for the entire dataset, and utilize the basis matrix W_0 as the novel representation and feed it into the data partition module, with the subscript "0" denoting the entire dataset.

It is worth noting that here we prefer using W_0 over typical label transformation methods such as CPLST [32] for the following reasons: 1) the proposed method does not rely on the true label matrix Y as in the formulation of CPLST, and 2) sample correlation can be explicitly embedded, which is suitable for data partition.

The data partition module takes as input W_0, and assigns a cluster label to each sample. According to this assignment, the visual feature matrix X and initial tag matrix D are reorganized into c sub-matrices denoted as $\{X_i\}_{i=1}^c \in \mathbb{R}^{n_i \times d}$ and $\{D_i\}_{i=1}^c \in \mathbb{R}^{n_i \times m}$ respectively, which are adopted for the establishment of local models later, see Section III-C. As mentioned earlier, n_i represents the number of samples contained in the i-th cluster, and thus we have $\sum_i n_i = n$.

Our approach makes no particular assumptions on the choice of partition algorithms, thus various methods can be considered, including k-means clustering, locality sensitive hashing (LSH) [33], and some adaptive methods such as Affinity Propagation clustering [34] or ISODATA [35], if sufficient prior knowledge is available. In our implementation, we use k-means clustering for its simplicity and efficiency.

C. Low-Rank Model Within Each Group

This section focuses on the construction of the local models for individual groups. As shown in (1), the j-th column in the

[1]The threshold for rare tags is 10, and the threshold for frequent tags is 300 for Corel5K, 1500 for IAPR TC12, and 1000 for Flickr30Concepts.

sparse coefficient matrix H_i can be interpreted as compressed k-dimensional representation for the j-th tag; and symmetrically, the l-th row in the basis matrix W_i can be considered as compressed representation for the l-th sample in the k-dimensional subspace. This new perspective reveals the inherent relations between the original spaces and the low-dimensional representation obtained by (1), and allows us to employ information obtained in the original spaces to improve the learning process of W_i and H_i.

Specifically, our method preserves local geometry structures in both the tag and image subspaces for each cluster. Similar to existing methods [29], [36], the proposed algorithm also assumes that the feature vector for each image can be linearly reconstructed by the feature vectors of several other images in the same cluster, thus the reconstruction coefficient matrix $S_i \in \mathbb{R}^{n_i \times n_i}$ can be obtained by

$$\begin{aligned} S_i^* &= \arg\min_{S_i} \left\{ \|X_i - S_i X_i\|_F^2 + \alpha \|S_i\|_1 \right\} \\ \text{s.t. } (S_i)_{ll} &= 0 \qquad \forall l \in 1, 2, \cdots, n_i \end{aligned} \tag{4}$$

where $(S_i)_{ll}$ denotes the l-th entry in the diagonal of S_i, and each row in $X_i \in \mathbb{R}^{n_i \times d}$ is a d-dimensional representation for an image.

Assume the tags for similar samples are also related, thus we have $Y_i \sim S_i Y_i$. According to the LLE [37] assumption, the structural information encoded in S_i should be robust to the sparse reconstruction process in (1), which means the low-dimensional representation of images can be equally reconstructed, *ie.*, $W_i \sim S_i W_i$. Thus R_{W_i} can be defined as

$$R_{W_i} = \|W_i - S_i W_i\|_F^2 . \tag{5}$$

Similarly, the reconstruction relationship in the original tag space can be obtained and passed to the low-dimensional representation as well. Denote the reconstruction coefficient matrix as T_i, which satisfies

$$\begin{aligned} T_i^* &= \arg\min_{T_i} \left\{ \|D_i - D_i T_i\|_F^2 + \mu \|T_i\|_1 \right\} \\ \text{s.t. } (T_i)_{jj} &= 0 \qquad \forall j \in 1, 2, \cdots, m \end{aligned} \tag{6}$$

where $(T_i)_{jj}$ denotes the j-th entry in the diagonal of T_i. The coefficient matrix T_i encodes the local geometry structures in the tag space, by assuming that the distribution of each tag can be linearly reconstructed by the distribution of other tags. Since H_i can be considered as the low-dimensional representation for tags, such relationship can be passed to H_i as well, thus we have $H_i \sim H_i T_i$.

Therefore, R_{H_i} can be written as below

$$R_{H_i} = \|H_i - H_i T_i\|_F^2 . \tag{7}$$

Note that by introducing the regularization specified in (5) and (7), the local geometry structures residing on D_i can be implicitly passed to Y_i. Therefore, consistency between tags and images are both maintained.

D. Global Consistency Among Local Models

As mentioned in Section I, optimizing each W_i and H_i independently for each cluster is not preferable due to potential overfitting, especially for the aforementioned cluttered clusters. Under such circumstances, images depicting the same concept may be partitioned into multiple clusters, whereas samples available for learning a specific model maybe insufficient. Therefore, the obtained local model is very likely to overfit the training data visible for the current cluster.

To fix this problem, a global consensus regularizer is introduced by assuming that each H_i is consistent with a global reference matrix H, as shown below

$$R_{g_i} = \|H - H_i\|_F^2 . \tag{8}$$

Recall that the columns of H_i can be considered as the k-dimensional tag representation in the i-th cluster, which should be consistent across different clusters. Therefore, the learning process for a cluttered cluster can be amended by forcing its tag representation H_i to be similar with the reference matrix H. In this way, the risk of overfitting could be alleviated by sharing information among images within various clusters.

The definition of (8) requires a reasonable initialization of H, the reference matrix. Fortunately, the coefficient matrix H_0 obtained in the pre-processing module offers a reasonable estimation for H, and thus can be adopted as its initialization.

IV. Optimization

The proposed formulation in (2) can be solved by many methods, such as basic gradient descent with line search. Here we present a method using coordinate descent [38], [39], which is easy to implement and converges fast.

A. Optimizing H_i

Define $G_i \triangleq \gamma(T_i - I)(T_i - I)^T$, $E_i \triangleq D_i^T W_i + \lambda H^T$ and $F_i \triangleq W_i^T W_i + \lambda I_k$, thus the objective function related to H_i is reduced as follows:

$$f(H_i) = tr\left\{H_i(-2E_i + H_i^T F_i + G_i H_i^T)\right\} + 2\beta \|H_i\|_1 . \tag{9}$$

Denote the entry in the tth row and jth column in H_i as $(H_i)_{tj}$, then the function related to $(H_i)_{tj}$ is a piece-wise parabolic function going upwards, which is convex and easy to obtain the optimal point

$$(H_i)_{tj} = \frac{\max\{(P_i)_{tj}, \beta\} + \min\{(P_i)_{tj}, -\beta\}}{(F_i)_{tt} + (G_i)_{jj}} \tag{10}$$

where

$$(P_i)_{tj} = (E_i)_{jt} - \sum_{\substack{r=1 \\ r \neq t}}^{k} (H_i)_{rj}(F_i)_{rt} - \sum_{\substack{r=1 \\ r \neq j}}^{m} (G_i)_{rj}(H_i)_{tr}. \tag{11}$$

B. Optimizing W_i

The optimal W_i can be obtained by a procedure similar to H_i. Define $Z_i \triangleq \eta(S_i - I)^T(S_i - I)$, $A_i \triangleq H_i D_i^T$ and $B_i \triangleq$

$H_i H_i^T$, thus we have

$$(W_i)_{lt} = \frac{(Q_i)_{lt}}{(B_i)_{tt} + (Z_i)_{ll}} \tag{12}$$

where

$$(Q_i)_{lt} = (A_i)_{tl} - \sum_{\substack{r=1 \\ r \neq t}}^{k} (B_i)_{rt}(W_i)_{lr} - \sum_{\substack{r=1 \\ r \neq l}}^{n_i} (W_i)_{rt}(Z_i)_{rl}. \tag{13}$$

C. Optimizing H

The updating rule of H can be readily obtained by taking the derivative of (2) and setting it to zero, as shown below

$$H = \frac{\sum_{i=1}^{c} H_i}{c}. \tag{14}$$

D. Implementation Issues

Some details are explained here. Firstly, when performing k-means clustering in the data-partition step, we empirically use correlation distance and set $c = 20$ for all the three datasets. In addition, a kNN strategy with 50 neighbors is adopted when calculating matrices S_i and T_i for the sake of efficiency. For S_i, the visual features and initial tags are concatenated as the feature for samples. The feature-sign method proposed in [40] is adopted when computing S_i and T_i, with $\alpha = 0.01$ and $\mu = 0.1$ for the local models, and $\alpha = 0.05$ and $\mu = 1$ for the pre-processing phase.

V. Experiments

Extensive empirical studies are carried out to validate the performance of the proposed approach. The experimental setup is outlined first, which is followed by the obtained results and the analysis of parameters.

A. Experimental Setup

We evaluate the proposed method on three datasets: the well-established benchmark datasets Corel5K and IAPR TC12, as well as a real-world dataset Flickr30Concepts. For each dataset, three types of features are used for evaluation, including the 1000-d SIFT BoW feature,[2] a 400-d composite feature obtained by merging 10 types of basic features,[3] and a 4096-d CNN feature extracted with a pre-trained 16-layer VGGNet [42], [43]. Statistics of all the three datasets are given in Table I.

For all the three datasets, 40% of the tags are randomly deleted, while ensuring that each image has at least one tag removed and one tag remained (hence samples associated with less than 2 tags are removed). Random deletions are performed 8 times and the averaged performance is reported. In addition, the evaluation method in [29] is adopted, as well as the same measurements: *Average Precision@N (AP@N)*, *Average Recall@N (AR@N)* and *Coverage@N (C@N)*.

Finally, to ensure a fair comparison, evaluations are performed only for the test set.

TABLE I
Statistics of Corel5K, Flickr30Concepts, and IAPR TC12. Counts of Tags Are Given in Format of "Mean/Maximum"

	Corel5K	Flickr30Concepts	IAPR TC12
Vocabulary Size	260	2,513	291
Nr. of Images	4,917	27,838	19,067
Tags per Image	3.4/5	8.3/70	5.7/23
Del. Tags per Image	1.4 (40%)	3.3 (40%)	2.2 (40%)
Test Set	491	2,807	1,898

B. Completion Results

To demonstrate the effectiveness of LSLR, we compare it with state-of-the-art annotation methods (MLR-GL [12], JEC [8], TagProp (σML) [9], and 2PKNN (ML) [10]) and several tag completion algorithms, namely, LR [26], Vote+ [24], Folksonomy [25], DLC [27], TMC [28], LSR [29] and its extended version DLSR [23].

Meanwhile, to further verify the rationality of the proposed method, we test it using ground truth tags for clustering, which is referred to as LSLR*g* (with the suffix *g* indicating *ground truth*). For LSLR*g*, we set $c = 50$ for Corel5K, $c = 40$ for IAPR TC12, and $c = 25$ for Flickr30Concepts (here c is determined according to the actual organization of images instead of cross validation). In particular, Flickr30Concepts is constructed by 30 semantic groups, thus we also test our method using these concepts, which is referred to as LSLR*d* (with the suffix *d* indicating *default*). Experimental results using the SIFT BoW feature are shown in Table II, and results using composite features are presented in Table III.

As seen from Tables II and III, TagProp achieves the best overall performance compared with other AIA methods, including 2PKNN. One possible reason is that, 2PKNN considers missing labels as negative ones and learns the model by maximizing the margin between positive and negative neighbors. Meanwhile, recent completion methods such as TMC and LSR outperform most AIA methods, due to their ability of exploiting information from initial labels. Finally, the proposed LSLR significantly outperforms previous methods for all the three datasets, which demonstrates the effectiveness of our strategy to approximate nonlinear reconstruction using a locality sensitive manner.

Furthermore, even better results are obtained by LSLR*g* and LSLR*d*, which is reasonable since these datasets are constructed with a number of semantic concepts, as with many other datasets. As a matter of fact, images are often organized by the semantic subsets they belonging to, both for personal photos and huge image depositories available on some social media. Being aware of such ubiquitous latent structures endows our method with enhanced capability to capture complex relatedness between samples and tags simultaneously.

[2]For Corel5K and IAPR TC12, their SIFT BoW features are downloaded from http://lear.inrialpes.fr/people/guillaumin/data.php., and the composite features are extracted by Lire [41]. For Flickr30Concepts [29], the two types of features provided by the authors are directly utilized.

[3]Including Color Correlogram, Color Layout, CEDD, Edge Histogram, FCTH, JCD, Jpeg Coefficient Histogram, RGB Color Histogram, Scalable Color, and SURF with Bag-of-Words model.

TABLE II
EVALUATION RESULTS USING 1000-D SIFT BOW FEATURE

	Corel5K			IAPR TC12			Flickr30Concepts		
	$AP@2$	$AR@2$	$C@2$	$AP@3$	$AR@3$	$C@3$	$AP@4$	$AR@4$	$C@4$
MLR-GL [12]	0.11	0.16	0.21	0.09	0.13	0.24	0.07	0.08	0.18
JEC [8]	0.14	0.21	0.27	0.15	0.21	0.34	0.07	0.07	0.15
TagProp (σML) [9]	0.19	0.27	0.34	0.18	0.27	0.43	0.06	0.07	0.18
2PKNN (ML) [10]	0.16	0.21	0.28	0.18	0.28	0.44	0.07	0.08	0.16
TMC [28]	0.23	0.33	0.40	0.14	0.20	0.37	0.19	0.21	0.37
DLC [27]	0.09	0.13	0.18	0.10	0.12	0.27	0.07	0.09	0.23
LSR [29]	0.28	0.42	0.50	–	–	–	0.30	0.36	0.60
DLSR [23]	0.28	0.42	0.50	0.23	0.31	0.55	0.31	0.37	0.61
LSLR	**0.37**	**0.54**	**0.65**	**0.30**	**0.47**	**0.67**	**0.38**	**0.47**	**0.75**
LSLR*g*	**0.39**	**0.58**	**0.69**	**0.33**	**0.51**	**0.71**	**0.41**	**0.52**	**0.81**
LSLR*d*	–	–	–	–	–	–	**0.42**	**0.54**	**0.83**

TABLE III
EVALUATION RESULTS USING COMPOSITE FEATURES

	Corel5K			IAPR TC12			Flickr30Concepts		
	$AP@2$	$AR@2$	$C@2$	$AP@3$	$AR@3$	$C@3$	$AP@4$	$AR@4$	$C@4$
MLR-GL [12]	0.12	0.19	0.24	0.09	0.13	0.24	0.22	0.28	0.47
JEC [8]	0.23	0.33	0.39	0.20	0.26	0.44	0.25	0.30	0.49
TagProp (σML) [9]	0.27	0.40	0.48	0.22	0.29	0.51	0.23	0.29	0.50
2PKNN (ML) [10]	0.23	0.33	0.40	0.20	0.28	0.43	0.26	0.30	0.46
Vote+ [24]	0.25	0.37	0.45	0.20	0.26	0.48	0.23	0.27	0.48
Folksonomy [25]	0.20	0.30	0.36	0.17	0.22	0.42	0.21	0.26	0.47
LR [26]	0.27	0.40	0.47	0.24	0.31	0.52	0.27	0.34	0.51
LSR [29]	0.33	0.48	0.58	0.30	0.41	0.64	0.37	0.45	0.67
DLSR [23]	0.34	0.50	0.59	0.30	0.41	0.65	0.38	0.48	0.71
LSLR	**0.39**	**0.56**	**0.65**	**0.33**	**0.50**	**0.70**	**0.40**	**0.50**	**0.76**
LSLR*g*	**0.41**	**0.60**	**0.71**	**0.36**	**0.55**	**0.75**	**0.43**	**0.54**	**0.82**
LSLR*d*	–	–	–	–	–	–	**0.45**	**0.56**	**0.84**

TABLE IV
EVALUATION RESULTS OF MULTIPLE VARIATIONS

	Corel5K			IAPR TC12			Flickr30Concepts		
	$AP@2$	$AR@2$	$C@2$	$AP@3$	$AR@3$	$C@3$	$AP@4$	$AR@4$	$C@4$
GLR-t	0.28	0.43	0.50	0.24	0.37	0.57	0.29	0.34	0.58
GLR-s (1000-d)	0.20	0.30	0.36	0.18	0.26	0.41	0.22	0.24	0.39
GLR-s (400-d)	0.25	0.37	0.45	0.22	0.33	0.49	0.25	0.30	0.51
GLR-s (4096-d)	0.32	0.49	0.58	0.28	0.43	0.63	0.36	0.45	0.68
LSLR-s (1000-d)	0.33	0.49	0.58	0.28	0.44	0.64	0.29	0.37	0.65
LSLR-s (400-d)	0.37	0.53	0.62	0.29	0.45	0.65	0.35	0.44	0.72
LSLR-s (4096-d)	0.39	0.57	0.67	0.31	0.48	0.69	0.40	0.51	0.78
LSLR-t (1000-d)	0.32	0.48	0.56	0.25	0.38	0.58	0.29	0.34	0.55
LSLR-t (400-d)	0.32	0.47	0.58	0.25	0.39	0.59	0.31	0.36	0.60
LSLR-t (4096-d)	0.36	0.53	0.62	0.29	0.45	0.66	0.32	0.38	0.62
LSLR (4096-d)	0.41	0.59	0.69	0.34	0.53	0.73	0.42	0.53	0.81
LSLR*g* (4096-d)	0.44	0.64	0.74	0.38	0.58	0.78	0.45	0.58	0.85

In addition, to investigate the respective effects of different visual features and regularization, we further evaluate several variations of the proposed method, including 1) GLR-t (Global Low-rank Reconstruction guided only by tag correlations with $c = 1$, $\lambda = 0$ and $\eta = 0$), 2) GLR-s (Global Low-rank Reconstruction guided only by visual similarity with $c = 1$, $\lambda = 0$ and $\gamma = 0$), 3) LSLR-t (Locality Sensitive Low-rank Reconstruction with $\eta = 0$), and 4) LSLR-s (Locality Sensitive Low-rank Reconstruction with $\gamma = 0$). The obtained results are summarized in Table IV, where different visual features are distinguished by their dimensions. Table IV also reports the evaluation results of standard LSLR and LSLR*g* using 4096-d CNN feature, as a complement to Tables II and III.

The first conclusion we can draw from Table IV is that, compared with their global counterparts, LSLR-t and LSLR-s models achieve significantly better performance, especially for the

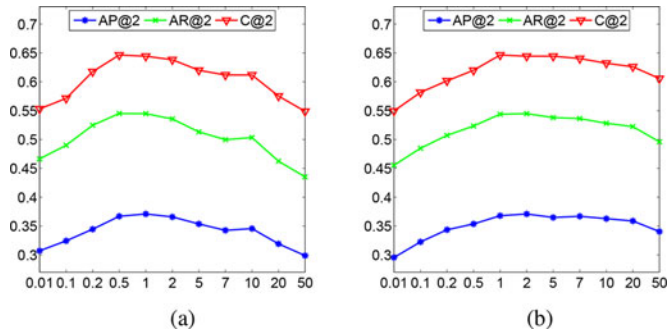

Fig. 2. Influences of η and γ on Corel5K using SIFT BoW feature.

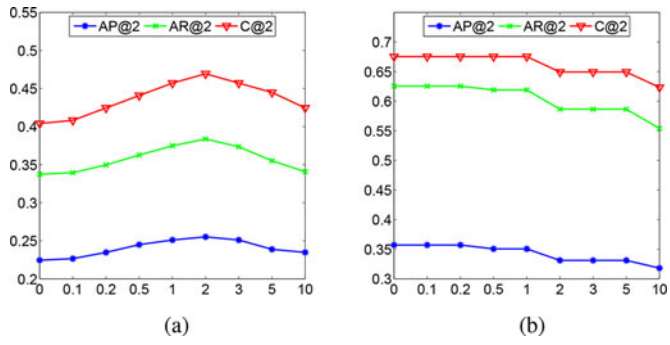

Fig. 3. Influences of λ on two clusters for Corel5K, by SIFT BoW feature. (a) λ for a cluttered cluster, (b) λ for a compact cluster.

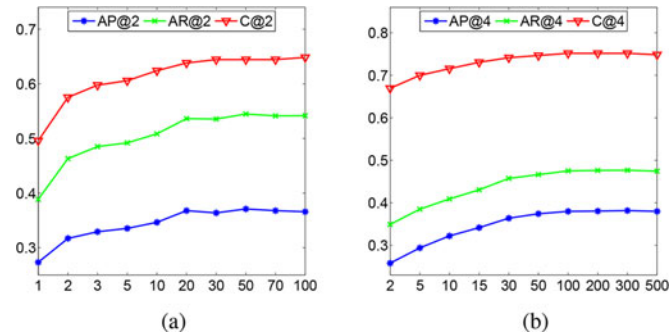

Fig. 4. Influences of basis number k, using SIFT BoW feature. (a) Corel5K. (b) Flickr30Concepts.

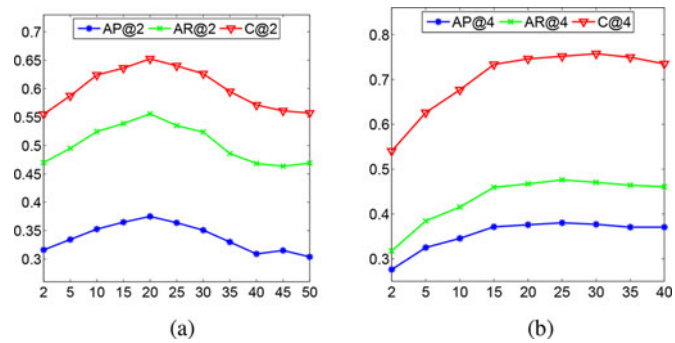

Fig. 5. Influences of cluster number c, using SIFT BoW feature. (a) Corel5K. (b) Flickr30Concepts.

latter ones with hand-engineered visual features. Table IV also allows us to analyze the relative contributions of visual features and initial tags. According to the results, both hand-engineered features yield inferior results to initial tags under the global setting, whereas they achieve comparable or better performance than tags in the locality sensitive models. CNN feature, on the other hand, consistently outperforms tag correlations under both the global and local settings. In particular, for standard LSLR, CNN feature yields substantially better performance compared with the results in Tables II and III.

C. Parameter Analysis

In this section, several main parameters are analyzed, including η, γ, λ and the basis number k. We empirically set an identical value of k for Corel5K and IAPR TC12, and only test its influences on Corel5K and Flickr30Concepts.

As shown in Fig. 2(a), the proposed method performs better as η gradually increases, then its performance begins to decline when increasingly larger values are used. The curve in Fig. 2(b) corresponding to γ exhibits a similar tendency.

Next, to examine the influence of λ, which controls the strength of our global consistency regularization, different values are tested on two clusters, one is a cluttered cluster, and the other one is a compact cluster containing initial tags including *bridge*, *arch*, *reflection* and *water*. As shown in Fig. 3(a), when $\lambda = 0$, the local model for the cluttered cluster is overfitted, leading to poor results. However, its performance gradually improves with λ growing larger, while the performance for the other cluster in Fig. 3(b) remains unchanged. This indicates that the model for the cluttered cluster is refined by global information, which justifies the necessity of introducing global consistency regularization. However, if λ becomes too large, the performance would degrade as well due to the loss of flexibility.

Meanwhile, since our method employs a matrix factorization scheme, it is necessary to specify an appropriate value for k, which is the number of columns in the basis matrix. As illustrated in Fig. 4, the proposed method can achieve fairly good performance with $k = 20$ for Corel5K (with 260 tags) and $k = 50$ for Flickr30Concepts (with 2513 tags). The impressive reduction of dimensions benefits from our locality sensitive strategy, which implicitly partitions tags into groups when performing clustering among samples.

Unless otherwise specified, all the evaluation results reported in this paper are obtained with the following parameters: $c = 20$, $k = 100$, $\eta = 1$, $\gamma = 2$, $\lambda = 0.5$, $\beta = 0.5$.

D. Analysis of Clustering

In this section, we analyze the influences of some parameters in clustering, including the number of clusters c and the sampling rate, denoted by ρ. Here the Corel5K dataset and Flickr30Concepts dataset are evaluated, since they are naturally organized by multiple semantic topics.

According to Fig. 5, the best cluster number are $c = 20$ for Corel5K and $c = 25$ for Flickr30Concepts, respectively. With

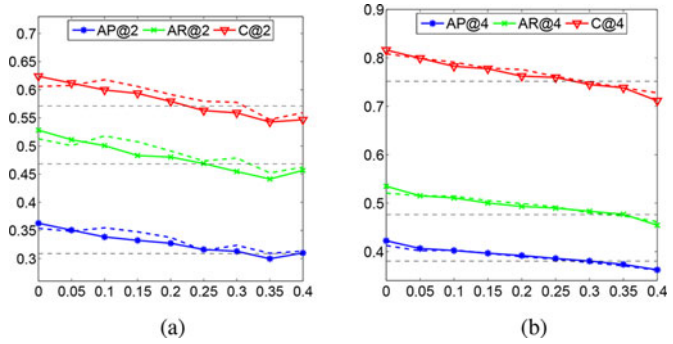

Fig. 6. Influences of sampling rate ρ, by SIFT BoW feature. The results obtained without pre-processing are shown in solid lines, results without the first step of pre-processing are shown in colored dash lines, and results of LSLR for $\rho = 0.4$ are shown in black dash lines. (a) Corel5K($c = 40$). (b) Flickr30Concepts($c = 25$).

smaller c, the advantages of locality sensitivity are weakened. However, as c becomes larger, the chance of encountering cluttered clusters would greatly increase, leading to significant performance deterioration. In practice, c can be estimated from prior knowledge. Take Flickr30Concepts as an example, given that its images are organized by 30 keywords and some images are associated with multiple concepts, an integer between 20 and 30 will be a reasonable estimate of c.

Another important factor is the sampling rate, which has been fixed to $\rho = 0.4$ previously. In this subsection, different values of ρ are examined for clustering (D used for (3) remains unchanged). In Fig. 6, the results obtained without pre-processing are shown in solid lines, and results obtained without the first step of pre-processing are shown in colored dash lines. In addition, our results obtained with $\rho = 0.4$ are also shown in black dash lines. According to Fig. 6, with the help of pre-processing, the Average Precision achieved by our method is equivalent to the case of $\rho = 0.32$ for Corel5K and $\rho = 0.3$ for Flickr30Concepts, which directly verifies the effectiveness of our pre-processing module. Meanwhile, without the removal of high-frequency and rare tags, the low-dimensional representation learnt in pre-processing leads to very limited improvement, which coincides with our analysis in Section III-B that these tags may hamper the learning of the new representation.

Finally, to give an intuitive impression on how could the proposed approach approximate the nonlinear reconstruction structure in a locality sensitive manner, several sample clusters are shown in Fig. 7, where samples associated with *water* (one of the most frequent tags in Corel5K) are partitioned into different semantic subsets. Note that the semantic ambiguity caused by this high-frequency tag is remarkably alleviated, which would definitely enhance the capability of our method, since recovering *beach* within the first cluster in Fig. 7 is much easier than recovering it from the entire dataset.

E. Analysis of Convergence and Efficiency

As indicated in Section IV, for each iteration of our alternating algorithm, the total computational complexity is $O\left(k^2 \times n + k \times \sum_{i=1}^{c} \left(nnz(H_i) + nnz(G_i) + nnz(Z_i)\right)\right)$,

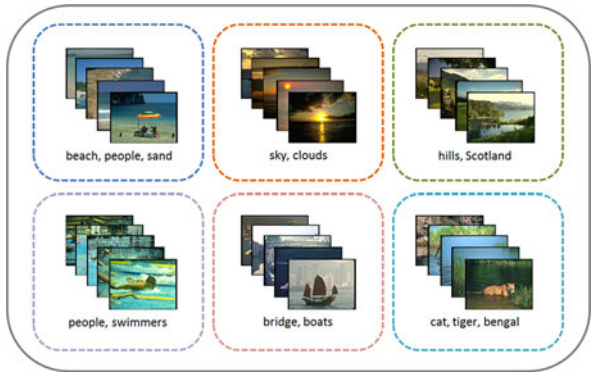

Fig. 7. Sample clusters related to high-frequency tag *water*, obtained from Corel5K. Other key words aside from *water* are shown beneath the images.

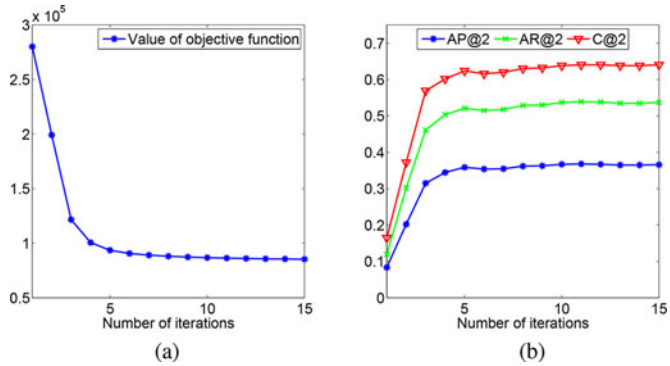

Fig. 8. Convergence properties of the optimization process on Corel5K, using SIFT BoW feature. (a) Function value. (b) Performance.

where $nnz(\cdot)$ denotes the number of non-zero entries in a sparse matrix. This enables our method to handle large datasets with its running time scales linearly with n.

Fig. 8 shows the convergence properties of the proposed method. According to Fig. 8(a), the proposed objective function converges within 10 iterations. In addition, the performance increases steadily as the iteration proceeds and finally reaches a satisfactory result, as shown in Fig. 8(b).

VI. Conclusion

In this paper we propose a locality sensitive low-rank model for image tag completion. The proposed method can capture complex correlations by approximating a nonlinear model with a collection of local linear models. To effectively integrate locality sensitivity and low-rank factorization, several adaptations are introduced, including the design of a pre-processing module and a global consensus regularizer. Our method achieves superior results on three datasets and outperforms pervious methods by a large margin.

References

[1] H.-F. Yu, P. Jain, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 593–601.

[2] M. M. Kalayeh, H. Idrees, and M. Shah, "NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 184–191.

[3] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2004, vol. 2, pp. 1002–1009.
[4] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.
[5] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, 2003, pp. 127–134.
[6] D. Putthividhy, H. T. Attias, and S. S. Nagarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3408–3415.
[7] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2006, vol. 2, pp. 2057–2063.
[8] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, vol. 5304, pp. 316–329.
[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 309–316.
[10] Y. Verma and C. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 836–849.
[11] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.
[12] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 2801–2808.
[13] Y. Verma and C. Jawahar, "Exploring SVM for image annotation in presence of confusing labels," in *Proc. Brit. Mach. Vis. Conf.*, 2013.
[14] H.-F. Yu, P. Jain, P. Kar, and I. S. Dhillon, "Large-scale multi-label learning with missing labels," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1307.5101.
[15] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1274–1282.
[16] S. Feng, Z. Feng, and R. Jin, "Learning to rank image tags with limited training examples," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1223–1234, Apr. 2015.
[17] Q. Wang, B. Shen, S. Wang, L. Li, and L. Si, "Binary codes embedding for fast image tagging with incomplete labels," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 425–439.
[18] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proc. Int. Conf. Multimedia*, 2010, pp. 491–500.
[19] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug. 2011.
[20] Y. Liu, F. Wu, Y. Zhang, J. Shao, and Y. Zhuang, "Tag clustering and refinement on semantic unity graph," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2011, pp. 417–426.
[21] K. Yang, X.-S. Hua, M. Wang, and H.-J. Zhang, "Tag tagging: Towards more descriptive keywords of image content," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 662–673, Aug. 2011.
[22] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 883–895, Jun. 2012.
[23] Z. Lin, G. Ding, M. Hu, Y. Lin, and S. S. Ge, "Image tag completion via dual-view linear sparse reconstructions," *Comput. Vis. Image Understand.*, vol. 124, pp. 42–60, 2014.
[24] B. Sigurbjörnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. Int. Conf. World Wide Web*, 2008, pp. 327–336.
[25] S. Lee, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Map-based image tag recommendation using a visual folksonomy," *Pattern Recog. Lett.*, vol. 31, no. 9, pp. 976–982, 2010.
[26] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. Int. Conf. Multimedia*, 2010, pp. 461–470.
[27] X. Liu, S. Yan, T.-S. Chua, and H. Jin, "Image label completion by pursuing contextual decomposability," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 8, no. 2, 2012, Art. ID 21.
[28] L. Wu, R. Jin, and A. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.
[29] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, "Image tag completion via image-specific and tag-specific linear sparse reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1618–1625.
[30] L. Ladicky and P. Torr, "Locally linear support vector machines," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 985–992.
[31] J. Lee, S. Bengio, S. Kim, G. Lebanon, and Y. Singer, "Local collaborative ranking," in *Proc. Int. Conf. on World Wide Web*, 2014, pp. 85–96.
[32] Y.-N. Chen and H.-T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Proc. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1529–1537.
[33] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, 2004, pp. 253–262.
[34] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
[35] G. H. D. J. H. Ball, *"ISODATA, a novel method of data analysis and pattern classification,"* Stanford Res. Inst., Menlo Park, CA, USA, Tech. Rep., 1965.
[36] X. Li, Y.-J. Zhang, B. Shen, and B.-D. Liu, "Low-rank image tag completion with dual reconstruction structure preserved," *Neurocomputing*, vol. 173, pp. 425–433, 2016.
[37] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.
[38] B.-D. Liu, Y.-X. Wang, Y.-J. Zhang, and B. Shen, "Learning dictionary on manifolds for image classification," *Pattern Recog.*, vol. 46, no. 7, pp. 1879–1890, 2013.
[39] B.-D. Liu, Y.-X. Wang, B. Shen, Y.-J. Zhang, and Y.-J. Wang, "Blockwise coordinate descent schemes for sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 5267–5271.
[40] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inform. Process. Syst.*, 2006, pp. 801–808.
[41] M. Lux and S. A. Chatzichristofis, "LIRe: Lucene image retrieval: An extensible Java CBIR library," in *Proc. Int. Conf. Multimedia*, 2008, pp. 1085–1088.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.
[43] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1408.5093.

Xue Li received the B.S. degree in electronic engineering from the Beijing Institute of Technology (BIT), Beijing, China, in 2011, and is currently working toward the Ph.D. degree in electronic engineering at Tsinghua University, Beijing, China.

Her research interests include image classification, automatic image annotation, and machine learning.

Bin Shen (S'09–M'16) received the B.S. and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2007 and 2009, respectively, and the Ph.D. degree in computer science from Purdue University, West Lafayette IN, USA, in 2014.

He is currently with Google Research, New York, NY, USA. His research interests include machine learning and its applications to data mining and computer vision.

Bao-Di Liu received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China.

He is currently an Assistant Professor with the College of Information and Control Engineering, China University of Petroleum, Qingdao, China. His research interests include computer vision and machine learning.

Yu-Jin Zhang (SM'99) received the Ph.D. degree in applied science from the Montefiore Institute, State University of Liege, Liege, Belgium, in 1989.

He was a Postdoctoral Fellow and Research Fellow with the Department of Applied Physics and the Department of Electrical Engineering, Delft University of Technology, Delft, the Netherlands, from 1989 to 1993. In 1993, he joined the Department of Electronic Engineering, Tsinghua University, Beijing, China, where he has been a Professor of Image Engineering since 1997. He has authored or coauthored more than 30 books and nearly 500 papers in the areas of image processing, image analysis, and image understanding.

Prof. Zhang is a Fellow of SPIE "for achievements in image engineering." He is the Director of the Academic Committee of the China Society of Image and Graphics.