Supervised Matrix Factorization Hashing for Cross-Modal Retrieval

Jun Tang, Ke Wang, and Ling Shao, *Senior Member,   IEEE*

Abstract— **The target of cross-modal hashing is to embed heterogeneous multimedia data into a common low-dimensional Hamming space, which plays a pivotal part in multimedia retrieval due to the emergence of big multimodal data. Recently, matrix factorization has achieved great success in cross-modal hashing. However, how to effectively use label information and local geometric structure is still a challenging problem for these approaches. To address this issue, we propose a cross-modal hashing method based on collective matrix factorization, which considers both the label consistency across different modalities and the local geometric consistency in each modality. These two elements are formulated as a graph Laplacian term in the objective function, leading to a substantial improvement on the discriminative power of latent semantic features obtained by collective matrix factorization. Moreover, the proposed method learns unified hash codes for different modalities of an instance to facilitate cross-modal search, and the objective function is solved using an iterative strategy. The experimental results on two benchmark data sets show the effectiveness of the proposed method and its superiority over state-of-the-art cross-modal hashing methods.**

Index Terms— **Cross-modal hashing, multimedia retrieval, collective matrix factorization, label consistency, local geometric consistency.**

I. INTRODUCTION

WITH the explosive growth of multimedia data, it is of particular interest to develop algorithms for scalable retrieval of similar visual content in large-scale datasets. In the community of computer vision, their applications can be found in various tasks such as visual categorization [1], [2], image/video retrieval [3], image annotation [4], object detection/recognition [5] and action recognition [6], [7]. Among them, the most fundamental scheme is the nearest neighbor (*NN*) search: given a query object, one should determine the most similar entity in a dataset and assign its label to the query. However, *NN* search is of linear time complexity with respect to the sample size of a dataset. To improve computational efficiency, tree based techniques and hashing techniques are the most popular methods nowadays. The weakness of tree based methods is that they are unable to effectively deal with high-dimensional data. By contrast, the main idea of hashing methods is to transform high-dimensional data into compact binary codes, and hence data dimension is not a significant problem for hashing based *NN* search. *NN* search is then conducted by finding the entities that have codes with a small Hamming distance from the query, which can be efficiently computed via fast bit-wise XOR operation. Additionally, the representation of binary codes can reduce tremendous storage cost compared with traditional feature descriptors.

Most previous hashing methods concentrated on producing binary codes for data within the same modality. They can be categorized as data-independent and data-dependent methods. One of the most well-known data-independent methods is locality sensitive hashing (LSH) [8], [9], which selects random linear projections independent of training data as hash functions. This idea was further explored in [10] and [11]. Data-dependent methods learns project functions from training data, which can produce more compact and discriminative hash codes than data-independent methods. One representative work is spectral hashing [12], in which hash codes are generated by thresholding the Laplace-Beltrami eigenfunctions of manifolds. PCA hashing [13] uses principle component analysis to learn hash functions. Moreover, various machine learning techniques have been applied to uni-modal hashing, such as manifold learning based approaches [14] and kernel learning based approaches [15], [16].

In practical applications, more and more multimodal data are generated owing to the rapid development of information industry. For example, a web image may have relevant text around it and a YouTube video often has associated tags and comments. Therefore, more research interest has been shifted from single-modal hashing to cross-modal hashing. It is worth noting that the idea of multi-view learning [17] is often used for cross-modal hashing. In terms of the utilization of label information, we can roughly group the existing cross-modal hashing methods into three categories, i.e., unsupervised methods [18]–[24], semi-supervised methods [25] and supervised methods [26]–[33]. In a general sense, label information is useful for learning more discriminative features despite that some labels may be corrupted and inaccurate [34], [35]. Therefore, supervised cross-modal hashing approaches can often achieve better retrieval performance.

J. Tang and K. Wang are with the Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230039, China (e-mail: tangjunahu@163.com; wangke4747@126.com).

L. Shao is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K. (e-mail: ling.shao@ieee.org).
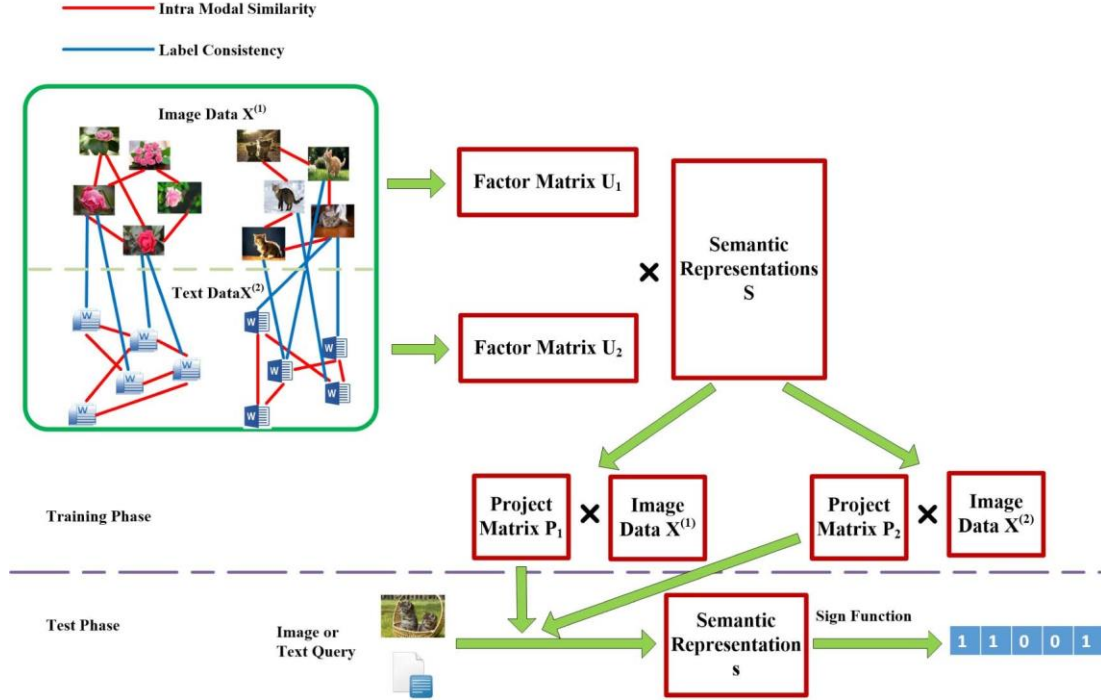
Fig. 1. Pipeline of the proposed approach.

Matrix factorization has a long history and many well-known techniques such as singular value decomposition, LU decomposition, QR decomposition and eigendecomposi- tion have been widely used in computer vision. Recent efforts are mainly devoted to investigating various constraints on the factors such as positivity, sparsity and rank. The most significant achievement brought by these developments may be the progress in compressive sensing. Matrix factorization also plays an important role in dealing with cross-modal hashing, which has achieved impressive results in mining semantic concepts or latent topics from image/text. Under the assumption that each modality of an instance has identical hash codes, Collective Matrix Factorization Hashing (CMFH) [19] learns unified hash codes via collective matrix factorization, which is used to find the latent factor model from different modalities. In [20], Latent Semantic Sparse Hashing (LSSH) learns latent semantic features for images and text using sparse coding and matrix factorization respectively, and then maps them into a joint abstract space to generate unified hash codes. Although the results of these two methods are promising, hash codes are learned in an unsupervised way and label information is not fully considered. Moreover, the preservation of intra-modal similarity is not taken into account.

To address these issues, we propose a supervised cross-modal hashing approach named Supervised Matrix Factoriza-tion Hashing (SMFH). We consider preserving intra-modal similarity as maintaining the local geometric structure of each modality, and enforce supervised learning by leveraging the label consistency between different modalities, which eventu-ally boil down to a graph regularization term in the objec-tive function. We use collective matrix factorization to learn a common latent semantic space, and incorporate the deduced

graph Laplacian to make data with the same class labels have similar semantic features. The given objective function is solved by alternate optimization. And the unified hash codes are generated from the semantic space by quantization. Fig. 1 depicts the working flow of the proposed algorithm.

Our work makes the following contributions:

✗ We put matrix factorization based cross-modal hashing into a supervised framework, which improves retrieval perfor-mance greatly.

✗ Our method not only considers the label information of data, but also explores the local manifold structure information in each modality, making the generated binary codes more discriminative.

✗ We conduct extensive experiments on two datasets to eval-uate the proposed SMFH. Experimental results demonstrate that our approach is competitive with state-of-the-art methods.

The rest of this paper is organized as follows. In Section 2, we briefly review some related work. Section 3 presents the detail of the proposed approach. In Section 4, we show experimental results in comparison to existing methods on two popular datasets, followed by the conclusion in Section 5.

II. RELATED WORK

We review the related work from four aspects.

The first category is unsupervised cross-modal hashing. The basic idea of learning cross-modal hashing functions is closely related to CCA [21], which projects data from two modalities into a common space where the correlation between them is maximized. Inter-media Hashing (IMH) [28] considers both inter-media consistency and intra-media con-sistency in learning hash functions, whose solution is based

on an integrated linear regression model. Linear Cross-Modal Hashing (LCMH) [18] borrows the idea from anchor graphs [36], which aims to preserve intra-similarity and inter-similarity in a common Hamming space. CMFH [19] is among the first to apply collective matrix factorization to cross-modal hashing, which learns latent factor model from different modalities and generates unified hash codes to support cross-modal search. Latent Semantic Sparse Hashing (LSSH) [20] also uses the assumption that the hash codes of different modalities of one instance are identical. It first uses sparse coding and matrix factorization to learn semantic features for images and text respectively, and then maps them into a joint abstract space to reduce semantic difference. And hash codes are generated in the obtained abstract space by quantization. LBMCH [37] simultaneously learns Hamming spaces for each modality and the semantic correspondences between these heterogeneous Hamming spaces. Full-space local topology extraction (FS-LTE) [23] first learns an intermediate common subspace by extracting the shared manifold structure across different modalities, and then obtains Hamming embedding by exploiting the local similarity in the intermediate subspace. Regularized cross-modal hashing (RCMH) [33] employs a three-step strategy, including hashing, regularization and partitioning, to learn a common multi-modal Hamming space. The last phase is designed to refine the decision planes obtained in the first phase.

The second category is semi-supervised cross-modal hashing. Semi-supervised hashing concentrates on using both labeled and unlabeled samples to learn hash functions. Although many algorithms for uni-modal data have been proposed [38], [39], this kind of work concerning multi-modal data is limited. Cheng et al. [25] presented a semi-supervised hashing framework using multi-graph learning. Semi-supervised information, which represents semantic similarity of each modality, is used to develop an efficient sequential learning algorithm for obtaining hash functions.

The third category is supervised cross-modal hashing. Cross-Modal Similarity Sensitive Hashing (CMSSH) [26] formulates learning hash codes as a binary classification problem and uses the boosting algorithm to tackle it. Cross-view Hashing (CVH) [27] extends spectral hashing [12] from single modality to multiple modalities. Zhen and Yeung [30] presented a generative model, i.e., the probabilistic latent factor model, to learn hash functions from multi-modal data. Co-Regularized Hashing (CRH) [30] uses a boosted co-regularized framework to learn a set of hash functions for each bit of the hash codes from one modality. Zhang and Li [31] proposed an efficient supervised multimodal hashing approach with linear time complexity, in which the adopted sequential learning strategy enables it fitting into large-scale data. Sparse multi-modal hashing (SM^{2}H) [22] models both intra-modality similarity and inter-modality similarity as a hypergraph and then learns the multi-modal dictionaries using hypergraph Laplacian sparse coding. Hash codes are generated by applying a sensitive Jaccard scheme to the obtained sparse representations. Semantics-Preserving Hashing (SePH) [32] transforms the predefined semantic affinities into a probability distribution and approximates it in Hamming space. Hash functions are learned using the kernel logistic regression with a sampling strategy.

The last one concerns methods that use geometric information to enhance representation discrimination. Zheng et al. [40] took into account the local manifold structure of data in learning sparse coding, making the obtained sparse representations better characterize the intrinsic geometric information of data. In [41], the geometric structures of the learned dictionary and sparse coefficients are simultaneously considered, which are described as a non-local self-similarity term and a graph Laplacian term in the objective function. Jiang et al. [42] addressed the efficient solution to graph regularized sparse coding. They decomposed the optimization of graph regularized sparse coding into two subproblems and alternately solved them until convergence.

Motivated by the promising results delivered by matrix factorization techniques in cross-modal retrieval [19], [20], we make further efforts to explore the application of collective matrix factorization by putting it into a supervised learning framework. Moreover, we study leveraging local geometric information to boost the discriminative power of the learned hash codes. The presented approach is developed by considering all these ingredients.

III. SUPERVISED MATRIX FACTORIZATION HASHING

Suppose that we have training data with two modalities, $\mathbf{X}^{(1)} = \{x_1^1, \cdots, x_n^1\}$ and $\mathbf{X}^{(2)} = \{x_1^2, \cdots, x_n^2\}$, which describe the same objects and n is the number of objects. Each object consists of an image and its accompanying text. For the i-th object, $x_i^{(1)} \in \mathsf{R}^{d_1}$ is a d_1-dimensional image descriptor, and $x_i^{(2)} \in \mathsf{R}^{d_2}$ is a d_2-dimensional text feature ($d_1 \neq d_2$ in most cases). Without loss of generality, we consider that the cross-modal data are zero-centered, i.e., $\sum_{i=1}^n x_i^{(1)} = 0$ and $\sum_{i=1}^n x_i^{(2)} = 0$.

The target of SMFH is to learn two groups of hash functions for image data and text data that are able to generate unified hash codes, i.e., $f(x^{(1)}) : \mathsf{R}^{d_1} \to \{-1, 1\}^k$ and $g(x^{(2)}) : \mathsf{R}^{d_2} \to \{-1, 1\}^k$, where k is the length of binary codes. To guarantee that each modality can yield identical hash codes, we first adopt the similar idea as that in [19] to project heterogeneous data into a common latent semantic space, while simultaneously considering the label information and the local manifold structure of each modality in acquiring the semantic space. Image data and text data represented by the obtained semantic features are then mapped into a Hamming space by quantization.

A. Collective Matrix Factorization

Most cross-modal hashing methods are built upon a reasonable assumption that heterogeneous data with the same semantic label share some common subspace [18], [27], [43], which is often named the latent semantic space. In the latent semantic space, the semantic representations of relevant data from different modalities are close to each other. We follow this idea due to its many successful applications.

In this work, we obtain the common semantic representation of heterogeneous data describing the same object via collective

matrix factorization [44], [45]. Collective matrix factorization was first used for relational learning, and one of its important applications is to predict unknown values of a relation when an entity dataset and multiple observed relations among entities are available.

Here we give a brief overview on collective matrix factorization. Suppose that we have two heterogeneous data matrices $\mathbf{X} \in \mathbb{R}^{d_1 \times n}$ and $\mathbf{Y} \in \mathbb{R}^{d_2 \times n}$, where n is the number of samples. The factors returned by collective matrix factorization are $\mathbf{U} \in \mathbb{R}^{d_1 \times k}$, $\mathbf{V} \in \mathbb{R}^{k \times n}$ and $\mathbf{Z} \in \mathbb{R}^{d_2 \times k}$, where k is the rank. $\mathbf{V}$ is the common factor in two reconstructions, i.e., $\mathbf{X} \approx f_1(\mathbf{UV})$ and $\mathbf{Y} \approx f_2(\mathbf{Z V})$, where f_i is a possibly-nonlinear function. The average decomposable losses are defined as:

$$L(\mathbf{U}, \mathbf{V}, \mathbf{Z} | \mathbf{X}, \mathbf{Y}) = \theta_1 L_1(\mathbf{U}, \mathbf{V} | \mathbf{X}) + \theta_2 L_2(\mathbf{Z}, \mathbf{V} | \mathbf{Y}), \quad (1)$$

where L_1 and L_2 are the decomposable loss function for $\mathbf{X} \approx f_1(\mathbf{U V})$ and $\mathbf{Y} \approx f_2(\mathbf{Z V})$ respectively, and $\sum_i \theta_i = 1$ are the balance parameters.

Collective matrix factorization provides a powerful facility to deal with multi-modal data. For the cross-modal hashing problem, we can use it to learn semantic features from different modalities:

$$O_{mf}(\mathbf{U}_1, \mathbf{U}_2, \mathbf{S}) = \alpha \| \mathbf{X}^{(1)} - \mathbf{U}_1 \mathbf{S} \|^2_F + (1 - \alpha) \| \mathbf{X}^{(2)} - \mathbf{U}_2 \mathbf{S} \|^2_F, \quad (2)$$

where $\mathbf{U}_1 \in \mathbb{R}^{d_1 \times k}$, $\mathbf{U}_2 \in \mathbb{R}^{d_2 \times k}$, $\mathbf{S} \in \mathbb{R}^{k \times n}$, and k is the number of latent semantic factors. The i-th column vector $\mathbf{s}_i$ of $\mathbf{S}$ is the common semantic representation of the i-th image data $x^{(1)}_i$ and its accompanying text. Here α is a balance parameter.

For out-of-sample instances, we map original image and text features into the latent semantic space by two linear projections respectively:

$$\mathbf{S}_I = \mathbf{P}_1 \mathbf{X}^{(1)}, \quad (3)$$
$$\mathbf{S}_T = \mathbf{P}_2 \mathbf{X}^{(2)}, \quad (4)$$

where $\mathbf{P}_1 \in \mathbb{R}^{k \times d_1}$ and $\mathbf{P}_2 \in \mathbb{R}^{k \times d_2}$.

Under the assumption that we have the same semantic representations for the mapped features, we minimize the following function to learn such two linear projections:

$$O_{lp}(\mathbf{P}_1, \mathbf{P}_2) = \| \mathbf{S} - \mathbf{S}_I \|^2_F + \| \mathbf{S} - \mathbf{S}_T \|^2_F$$
$$= \| \mathbf{S} - \mathbf{P}_1 \mathbf{X}^{(1)} \|^2_F + \| \mathbf{S} - \mathbf{P}_2 \mathbf{X}^{(2)} \|^2_F. \quad (5)$$

B. Mixed Graph Regularization Term

In this part, we describe the formulation of using label information and local geometric structure.

To make use of label information, we first model the label consistency between the image modality and the text modality, and define the semantic affinity metric of two data $x^{(1)}_i$ and $x^{(2)}_j$ from different modalities as follows:

$$c_{ij} = \begin{cases} 1, & if \ x^{(1)}_i \ and \ x^{(2)}_j \ have \ the \ same \ category \\ 0. & otherwise \end{cases} \quad (6)$$

We then define the local similarity metric to model the intra-modal similarity in each modality. The intra-modal similarity $s^{(m)}_{ij}$ $(m = 1, 2)$ of two data $x^{(m)}_i$ and $x^{(m)}_j$ from the same modality is defined as follows:

$$s^{(m)}_{ij} = \begin{cases} 1, & if \ x^{(m)}_i \in \mathbf{N}_k(x^{(m)}_j) \ or \ x^{(m)}_j \in \mathbf{N}_k(x^{(m)}_i) \\ 0, & otherwise \end{cases} \quad (7)$$

where $\mathbf{N}_k(\cdot)$ denotes the set of k-nearest neighbors. According to our empirical study, we set the number of nearest neighbors to 5 throughout this paper.

To maintain the label consistency between two modalities and preserve the intra-modal similarity of each modality in the common semantic space, we can minimize the following function:

$$O_{sc}(\mathbf{S}) = \frac{1}{2} \left(\sum_{m=1}^{2} \sum_{i,j=1}^{n} s^{(m)}_{ij} \| \mathbf{s}_i - \mathbf{s}_j \|^2 + \sum_{i,j=1}^{n} c_{ij} \| \mathbf{s}_i - \mathbf{s}_j \|^2 \right)$$
$$= \frac{1}{2} \sum_{i,j=1}^{n} (s^{(1)}_{ij} + s^{(2)}_{ij} + c_{ij}) \| \mathbf{s}_i - \mathbf{s}_j \|^2$$
$$= \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \| \mathbf{s}_i - \mathbf{s}_j \|^2, \quad (8)$$

where $w_{ij} = s^{(1)}_{ij} + s^{(2)}_{ij} + c_{ij}$.

By some algebraic calculation, the function in Eq. (8) can be reformulated as:

$$O_{sc}(\mathbf{S}) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \| \mathbf{s}_i - \mathbf{s}_j \|^2$$
$$= tr(\mathbf{S}(\mathbf{D} - \mathbf{W})\mathbf{S}^T)$$
$$= tr(\mathbf{S}\mathbf{L}\mathbf{S}^T), \quad (9)$$

where $\mathbf{W} \in \mathbb{R}^{n \times n}$ consists of w_{ij} in the i-th row and the j-th column. $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose entries are the column sum of $\mathbf{W}$, i.e., $\mathbf{D}_{ii} = \sum_j w_{ij}$. Hence $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is with the form of Laplacian matrix. So far we have obtained a mixed graph regularization term describing both label information and local manifold structure.

C. Overall Objective Function

The overall objective function, consisting of the collective matrix factorization term O_{mf} in Eq. (2), the linear projection term O_{lp} in Eq. (5), the mixed graph regularization term O_{sc} in Eq. (9) and a Frobenius term, is given as follows:

$$\min_{\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{S}} O(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{S})$$
$$= O_{mf} + O_{lp} + O_{sc} + \lambda R(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{S})$$
$$= \alpha \| \mathbf{X}^{(1)} - \mathbf{U}_1 \mathbf{S} \|^2_F + (1 - \alpha) \| \mathbf{X}^{(2)} - \mathbf{U}_2 \mathbf{S} \|^2_F$$
$$+ \beta (\| \mathbf{S} - \mathbf{P}_1 \mathbf{X}^{(1)} \|^2_F + \| \mathbf{S} - \mathbf{P}_2 \mathbf{X}^{(2)} \|^2_F)$$
$$+ \gamma \, tr(\mathbf{S}\mathbf{L}\mathbf{S}^T) + \lambda R(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{S}), \quad (10)$$

where the regularization term is defined as $R(\cdot) = \| \cdot \|^2$ to resist overfitting. β, γ and λ are the tradeoff parameters of the corresponding terms. We set $\alpha = 0.5, \beta = 100, \gamma = 1$ and $\lambda = 0.01$ throughout this paper.

The optimization problem stated in Eq. (10) is hard to be directly solved due to its non-convexity with five matrix variables $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2$ and $\mathbf{S}$. However, it is convex with respect to any one of the five variables in the case that the

Algorithm 1 Supervised Matrix Factorization Hashing

Require: Image feature matrix $\mathbf{X}^{(1)}$ and text feature matrix $\mathbf{X}^{(2)}$, the semantic labels $\bar{L}$ of bimodal data, parameters $\alpha, \beta, \gamma, \lambda$ and the length of hash codes k.

Ensure: Unified hash codes $\mathbf{H}$, projection matrix $\mathbf{P}_1, \mathbf{P}_2$.

1: Center $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ by means, and construct the Laplacian matrix $\mathbf{L}$ according to Eqs. (8) and (9).

2: Initialize $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{S}$ by random matrices respectively.

3: **repeat**

4: Fix $\mathbf{P}_1, \mathbf{P}_2, \mathbf{S}$, update $\mathbf{U}_1, \mathbf{U}_2$ by Eq. (11) and Eq. (12) respectively.

5: Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2$, update $\mathbf{S}$ by solving the Sylvester equation in Eq. (13).

6: Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{S}$, update $\mathbf{P}_1, \mathbf{P}_2$ by Eq. (14) and Eq. (15) respectively.

7: **until** convergence.

8: $\mathbf{H} = sign(\mathbf{S})$

---

remaining four are treated as constants. Therefore, we can adopt an iterative strategy to solve this optimization problem following the listed three steps until convergence.

Step 1: Fix $\mathbf{P}_1, \mathbf{P}_2, \mathbf{S}$, let $\frac{\partial \mathbf{O}}{\partial \mathbf{U}_1} = 0, \frac{\partial \mathbf{O}}{\partial \mathbf{U}_2} = 0$, the obtain:

$$\mathbf{U}_1 = \mathbf{X}^{(1)}\mathbf{S}^T (\mathbf{S}\mathbf{S}^T + \frac{\lambda}{\alpha}\mathbf{I})^{-1}, \qquad (11)$$

$$\mathbf{U}_2 = \mathbf{X}^{(2)}\mathbf{S}^T (\mathbf{S}\mathbf{S}^T + \frac{\lambda}{1-\alpha}\mathbf{I})^{-1}, \qquad (12)$$

where $\mathbf{I}$ is the identity matrix.

Step 2: Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2$, let $\frac{\partial \mathbf{O}}{\partial \mathbf{S}} = 0$, then obtain:

$$\mathbf{A}\mathbf{S} + \mathbf{S}\mathbf{B} + \mathbf{C} = 0, \qquad (13)$$

where

$\mathbf{A} = 2(\alpha\mathbf{U}^{(1)^T}\mathbf{U}^{(1)} + (1-\alpha)\mathbf{U}^{(2)^T}\mathbf{U}^{(2)} + (2\beta + \lambda)\mathbf{I}),$

$\mathbf{B} = \mathbf{L} + \mathbf{L}^T,$

$\mathbf{C} = -2(\alpha\mathbf{U}_1^T\mathbf{X}^{(1)} + (1-\alpha)\mathbf{U}_2^T\mathbf{X}^{(2)} + \beta(\mathbf{P}_1\mathbf{X}^{(1)} + \mathbf{P}_2\mathbf{X}^{(2)})).$

Note that Eq. (13) is a Sylvester equation [46]. We can solve it using the lyap function of MATLAB.

Step 3: Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{S}$, let $\frac{\partial \mathbf{O}}{\partial \mathbf{P}_1} = 0, \frac{\partial \mathbf{O}}{\partial \mathbf{P}_2} = 0$, the obtain:

$$\mathbf{P}_1 = \mathbf{S}\mathbf{X}^{(1)^T} (\mathbf{X}^{(1)}\mathbf{X}^{(1)^T} + \frac{\lambda}{\beta}\mathbf{I})^{-1}, \qquad (14)$$

$$\mathbf{P}_2 = \mathbf{S}\mathbf{X}^{(2)^T} (\mathbf{X}^{(2)}\mathbf{X}^{(2)^T} + \frac{\lambda}{\beta}\mathbf{I})^{-1}. \qquad (15)$$

The SMFH approach is summarized in Algorithm 1. When a new query x_q comes, SMFH firstly generates the semantic representation s_q according to Eq. (3) or Eq. (4). And then we can obtain the hash codes using the sign function $\mathbf{H}^{x_q} = sign(s_q)$.

IV. EXPERIMENTS

In this section, we evaluate the proposed approach on two benchmark datasets, i.e., the Wiki dataset [47] and the NUS-WIDE dataset [48]. To validate the proposed method,

we comprehensively compare it with some state-of-the-art cross-modal hashing approaches, including LCMH [18], CCA [21], IMH [28], LSSH [20], CMFH [19], CVH [27], SCM_orth [31] and SCM_seq [31].

A. Experimental Setup

The details of the evaluated datasets are shown in Table II. Wiki. The Wiki dataset [47] is collected from "Wikipedia featured articles." This dataset consists of 2866 multimedia documents grouped into 10 semantic categories and each document is an image-text pair. Each image is represented by a 128-dimensional Bag-of-Visual-Words SIFT feature vector and each text is represented by a 10-dimensional topics vector. Following the experimental protocol in [19], [20], and [47], we randomly select 2173 documents as the training set and the rest as the test set.

NUS-WIDE. The NUS-WIDE dataset [48] is a web image dataset downloaded from Flickr, containing about 270K images with annotated tags from 81 semantic concepts. Each image with its tag annotations can be taken as an image-text pair. Following [49], we choose 10 largest categories consisting of 186577 images as the experimental data. The selected 10 categories include animal, buildings, clouds, grass, and lake, etc. Each image is represented by a 500-dimensional Bag-of-Visual-Words SIFT histogram and each text is represented by a Bag-of-Words feature vector based on the top 1000 most frequent tags. In the chosen experimental data, we randomly sample 1866 image-text pairs as the test set and 5000 image-text pairs randomly selected from the remaining documents are used as the training set.

We perform two kinds of cross-modal retrieval tasks. One is 'Img to Txt', i.e., using image queries to search relevant text. The other is 'Txt to Img', i.e., using text queries to search relevant images. We employ the mean Average Precision (mAP) to evaluate the performance of the cross-modal retrieval results:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP(q_i),$$

where q_i is a query and N is the number of queries. $AP(\cdot)$ is computed by:

$$AP(q) = \frac{1}{T} \sum_{r=1}^{R} P_q(r)\xi(r),$$

where T is the number of relevant entities in the retrieved set. $P_q(r)$ represents the precision of top r retrieved entities and $\xi(r)$ is defined as an indicator function whose value is 1 if the r th entity is relevant to query or 0 otherwise.

Moreover, we study two types of performance curves on the Wiki dataset, i.e., the *precision-recall* curve and the *topN-precision* curve. The *precision-recall* curve shows the precision as a function of different recall, which is a frequently used measurement on information retrieval performance. The *topN-precision* curve demonstrates the variation of precision with respect to different numbers of retrieved samples. As the evaluated data are randomly selected, all the reported results are averaged over 10 independent trials.
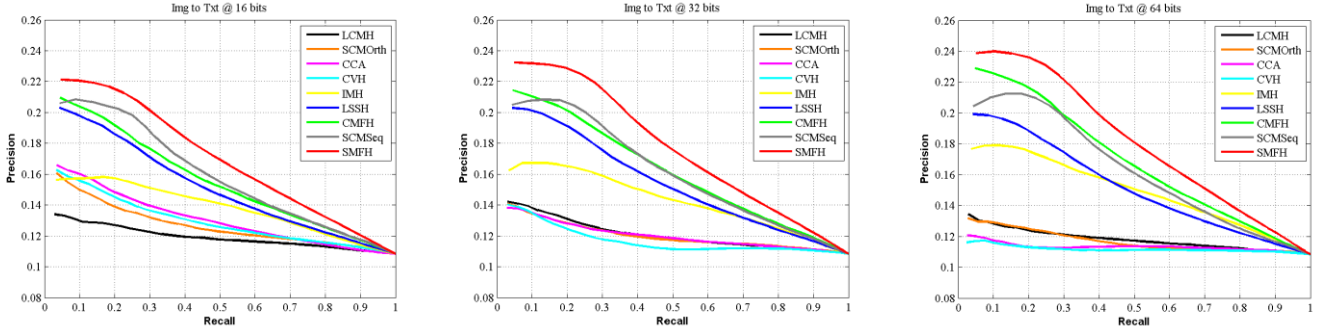
Fig. 2.   PR-Curves on the Wiki dataset Varying Code Length (Img to   Txt).
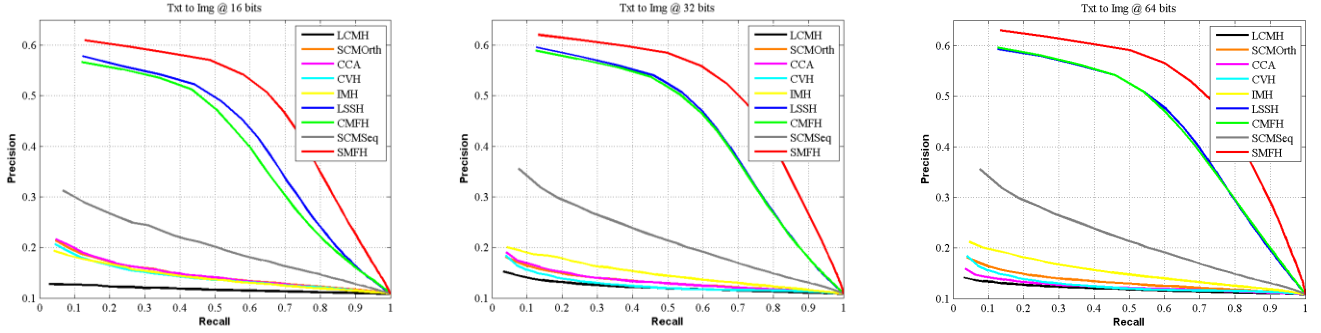

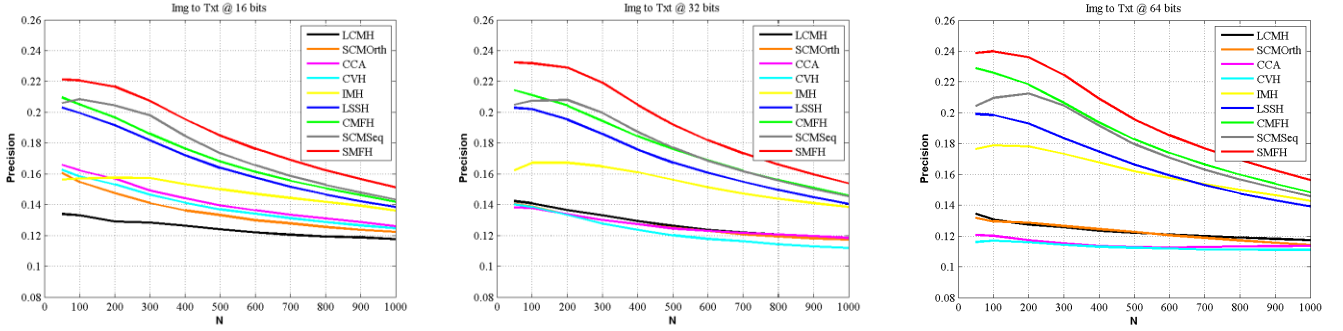Fig. 3.   PR-Curves on the Wiki dataset Varying Code Length (Txt to   Img).

Fig. 4.   topN-Precision Curves on the Wiki dataset Varying Code Length (Img to   Txt).

B. Experimental Results

We first study the experimental results on the Wiki dataset. Table I reports the mAP scores of different methods with respect to different hash code lengths. We have two observations from Table I. Firstly, our SMFH approach yields much better mAP results than the state-of-the-art methods in the overall experiments, which demonstrates its superiority over the compared methods. Specifically, SMFH outperforms CMFH by about 20% in the case that the hash code length is 16 bits, which demonstrates that SMFH makes a substantial improvement on CMFH. Secondly, with the increment of the hash code length, the performance of SMFH continuously increases, which can be attributed to its ability to better preserve the label consistency across modalities and the local geometric information with longer codes. However, the performance of some compared methods such as   CCA and CVH degrades to some extent, which is also observed   in

[19] and [32]. As discussed in [19] and [32], it can be boiled down to that the eigendecomposition solution may make the lower bits indiscriminative.

The *precision-recall* curves are plotted in Fig. 2 and Fig. 3, and the *topN-precision* curves are plotted in Fig. 4 and Fig. 5. We have three observations from these experimental results. Firstly, SMFH outperforms all the compared methods, which is consistent with the mAP results. Secondly, the behavior of LCMH and CCA looks like random guess and the results are almost meaningless. Thirdly, even with relatively short binary codes (16 bits), SMFH still defeats the compared methods with longer binary codes (32 bits), showing the advantage of storage efficiency of our approach.

We then study the results on the NUS-WIDE dataset, which are detailed in Table III. The results are somewhat different from those on the Wiki dataset. SMFH outperforms most of the compared methods apart from SCM_seq, which is   consistent
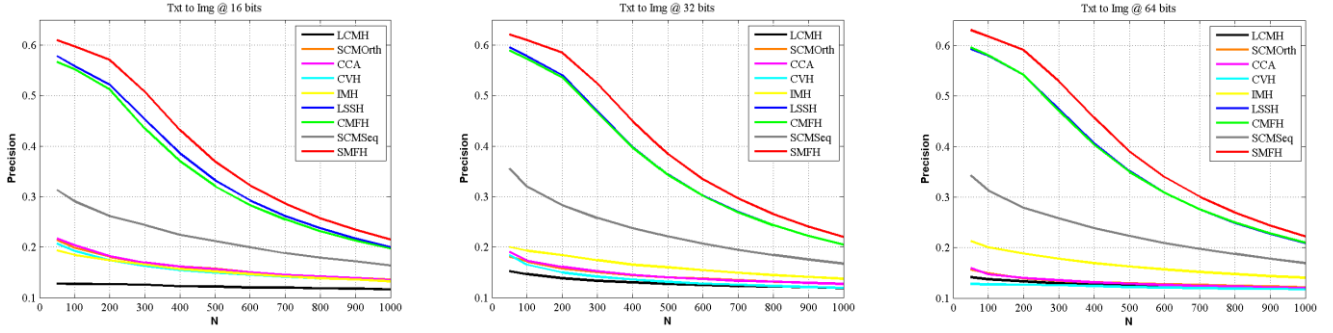
Fig. 5.   topN-Precision Curves on the Wiki dataset Varying Code Length (Txt to Img).

TABLE I

mAP RESULTS ON THE WIKI DATASET. ITEMS IN BOLD INDICATE THE BEST PERFORMANCE

Task	Methods	Hash Code Length			
		16bits	32bits	64bits	128bits
Img to Txt	LCMH	0.1551	0.1490	0.1352	0.1324
	CCA	0.1714	0.1547	0.1465	0.1286
	IMH	0.1952	0.2003	0.2084	0.2097
	LSSH	0.2162	0.2248	0.2255	0.2190
	CMFH	0.2116	0.2260	0.2394	0.2414
	CVH	0.1718	0.1573	0.1508	0.1351
	SCM_orth	0.1551	0.1406	0.1375	0.1247
	SCM_seq	0.2341	0.2410	0.2455	0.2563
	SMFH	**0.2572**	**0.2759**	**0.2863**	**0.2913**
Txt to Img	LCMH	0.1229	0.1292	0.1252	0.1258
	CCA	0.1590	0.1415	0.1300	0.1151
	IMH	0.1508	0.1581	0.1636	0.1668
	LSSH	0.5010	0.5246	0.5324	0.5387
	CMFH	0.4830	0.5235	0.5361	0.5318
	CVH	0.1505	0.1347	0.1214	0.1136
	SCM_orth	0.1554	0.1398	0.1293	0.1137
	SCM_seq	0.2257	0.2459	0.2437	0.2476
	SMFH	**0.5784**	**0.6040**	**0.6163**	**0.6219**

with the experiments on the Wiki dataset. Additionally, we can observe that SCM_seq outperforms SMFH in the first task while producing quite close results compared with SMFH in the second task. We conjecture that it is because the identical constraint of hash codes from each modality in CMFH and SMFH is too strict for the NUS-WIDE dataset. However, in consideration of the huge advantage of SMFH in the previous experiments, we can learn that SMFH is competitive with the compared methods.

Finally, we investigate the training time of all these methods. The experiments are conducted on the NUS-WIDE datset and run on a PC with Intel (R) Core (TM) CPU i7-4790@3.60GHz and 16GB RAM. Here we only evaluate the case that the code length is 16 bits. The results are reported in Table IV. We can observe that the time consumption of SMFH is of the same order of magnitude as that of CMFH. Although both of them involve the computation of matrix inversion, the time cost is acceptable in comparison with that of LSSH. As SMFH needs to solve the Sylvester equation in an iterative way, it spends a little more time than CMFH in the training phase.

C. Effect of Graph Laplacian Term

To validate the advantage of the mixed graph Lapla-cian term, we give two variants of the proposed algorithm for comparison. The first one only uses local geometric

TABLE II

THE DETAILS OF THE EVALUATED DATASETS

Datasets	Wiki	NUS-WIDE
Dataset Size	2866	186577
Training Set Size	2173	5000
Test Set Size	693	1866
Num. of Categories	10	10

consistency to construct the graph Laplacian term,and we redefine $w_{ij} = s_{ij}^{(1)} + s_{ij}^{(2)}$ in Eq. (8). The second one only uses label consistency to build the graph Laplacian term, where $w_{ij} = c_{ij}$. These two methods are referred to as SMFH_lc and SMFH_gc for convenience. We evaluate their performance in the case of 16 bits length. The mAP results are summarized in Table V. We can learn that the graph Laplacian term, describing label consistency or local geometric consistency, produces a positive result. Notably, the mixed graph Laplacian term leads to better performance than the individual one, which verifies the effectiveness of the proposed strategy.

D. Parameter Sensitivity

There are four parameters in our SMFH, including α, β, γ and λ in the objective function. In the previous experiments,

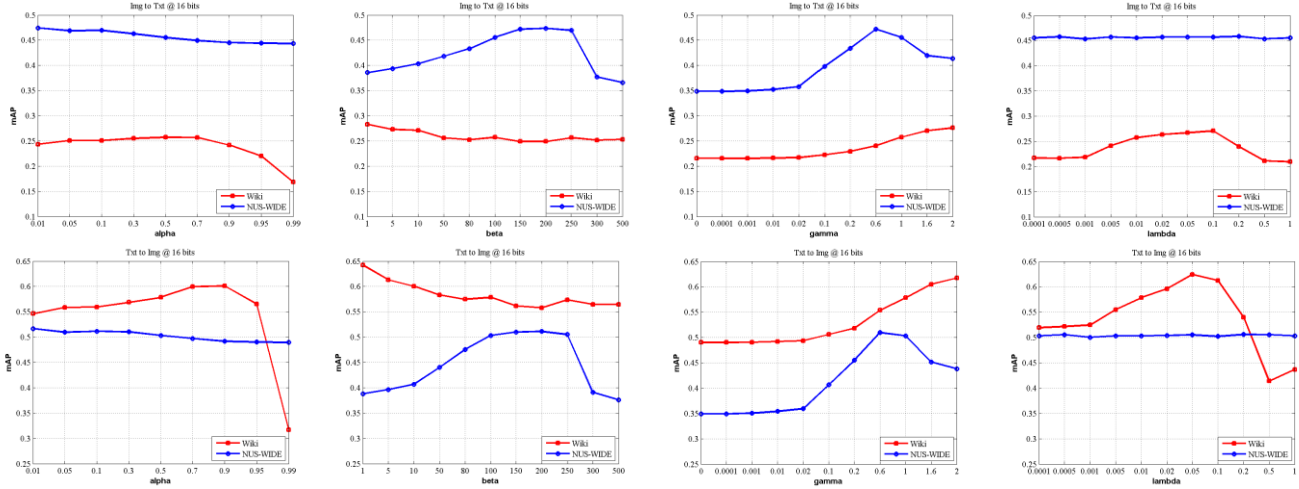| Task	Methods	Hash Code Length			
		16bits	32bits	64bits	128bits
Img to Txt	LCMH	0.3335	0.3293	0.3249	0.3218
	CCA	0.3606	0.3538	0.3499	0.3472
	IMH	0.4029	0.4090	0.4268	0.4248
	LSSH	0.3838	0.3837	0.3903	0.3864
	CMFH	0.3728	0.3793	0.3788	0.3771
	CVH	0.3738	0.3627	0.3555	0.3506
	SCM_orth	0.3907	0.3758	0.3646	0.3542
	SCM_seq	**0.5219**	**0.5336**	**0.5365**	**0.5189**
	SMFH	0.4553	0.4623	0.4658	0.4680
Txt to Img	LCMH	0.3493	0.3400	0.3249	0.3218
	CCA	0.3598	0.3533	0.3502	0.3482
	IMH	0.4002	0.4091	0.4315	0.4317
	LSSH	0.4150	0.4114	0.4165	0.4132
	CMFH	0.3734	0.3785	0.3823	0.3819
	CVH	0.3757	0.3636	0.3567	0.3526
	SCM_orth	0.3875	0.3696	0.3608	0.3538
	SCM_seq	**0.5090**	**0.5219**	**0.5297**	0.5040
	SMFH	0.5033	0.5056	0.5065	**0.5079**

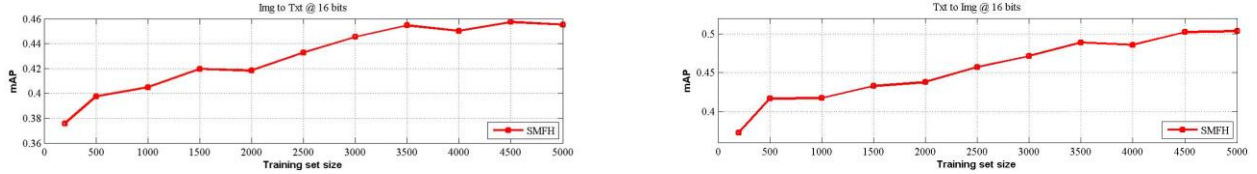Fig. 6.   Parameter Sensitivity Analysis.

Fig. 7.   Effect of Training Set Size on mAP on the NUS-WIDE   dataset.

we empirically set $\alpha = 0.5$, $\beta = 100$, $\gamma = 1$ and $\lambda = 0.01$. Here we conduct experiments to analyze the effects of different parameter settings on the algorithm performance. The experiments are performed by varying the value of one parameter while fixing the other three, in which the code length is fixed to 16 bits. And the mAP results are used to evaluate the performance variation with respect to different parameter values. Fig. 6 plots the results on the test datasets. We can learn that all the parameters are not sensitive and SMFH can yield satisfactory results in a wide range of parameter   values.

E. Training Set Size

In this part, we analyze the effect of varying training set size on the NUS-WIDE dataset. The code length is also fixed to 16 bits in this evaluation. We vary the training set   size from 200 to 5000 with the gap of 500. Fig. 7 shows the mAP results of our SMFH as a function of different training set sizes. We can observe that the mAP results increase quickly as the training set size grows from 200 to 3500. When the training set size is larger than 3500, the performance of SMFH tends to converge. This demonstrates that SMFH is suitable for

TABLE IV
TRAINING TIME OF ALL THE EVALUATED METHODS ON
THE NUS-WIDE DATASET

Methods	Time(s)
LCMH	25.2
CCA	0.2
IMH	67.7
LSSH	2716.0
CMFH	305.8
CVH	1.9
SCM_orth	0.8
SCM_seq	8.2
SMFH	312.0

TABLE V
EFFECT OF DIFFERENT GRAPH LAPLACIAN TERMS

	Methods	Wiki	NUS-WIDE
Img to Txt	CMFH	0.2116	0.3728
	SMFH_lc	0.2164	0.3816
	SMFH_gc	0.2519	0.4153
	SMFH	0.2572	0.4553
Txt to Img	CMFH	0.4830	0.3734
	SMFH_lc	0.4917	0.3816
	SMFH_gc	0.5697	0.4434
	SMFH	0.5784	0.5033

large-scale datasets and can achieve satisfactory performance with a reasonably small training set.

V. CONCLUSION

In this paper, we have proposed an effective cross-modal hashing approach named Supervised Matrix Factorization Hashing. The proposed approach uses collective matrix factorization to generate unified hash codes to achieve cross-modal search. In particular, we use both label consistency across different modalities and local geometric consistency in each modality to make the learned hash codes more discriminative. These two ingredients are formulated as a mixed graph Laplacian regularization term in the objective function. Experimental results on two benchmark datasets have demonstrated that SMFH outperforms several state-of-the-art methods.

REFERENCES

[1] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multiview matrix completion for multilabel image classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2355–2368, Aug. 2015.

[2] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, Aug. 2014.

[3] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 956–966, Mar. 2015.

[4] C. Xu, T. Liu, D. Tao, and C. Xu, "Local Rademacher complexity for multi-label learning," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1495–1507, Mar. 2016.

[5] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015.

[6] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB-D action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2015.2491925.

[7] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, pp. 1–15, 2015, doi: 10.1007/s11263-015-0861-6.

[8] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proc. VLDB*, 1999, pp. 518–529.

[9] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Proc. 47th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2006, pp. 459–468.

[10] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1509–1517.

[11] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1092–1104, Jun. 2012.

[12] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1753–1760.

[13] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3424–3431.

[14] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1–8.

[15] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2074–2081.

[16] S. Wang, Q. Huang, S. Jiang, and Q. Tian, "Scalable semi-supervised multiple kernel learning for real-world image applications," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1259–1274, Aug. 2012.

[17] C. Xu, D. Tao, and C. Xu, "Multi-view intact space learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2531–2544, Dec. 2015.

[18] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 143–152.

[19] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2083–2090.

[20] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 415–424.

[21] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 817–824.

[22] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[23] L. Zhang, Y. Zhang, R. Hong, and Q. Tian, "Full-space local topology extraction for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2212–2224, Jul. 2015.

[24] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.

[25] J. Cheng, C. Leng, P. Li, M. Wang, and H. Lu, "Semi-supervised multi-graph hashing for scalable similarity search," *Comput. Vis. Image Understand.*, vol. 124, pp. 12–21, Jul. 2014.

[26] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3594–3601.

[27] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc.-Int. Joint Conf. Artif. Intell.*, 2011, vol. 22. no. 1, p. 1360.

[28] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[29] X. Chen and D. Cai, "Large scale spectral clustering via landmark-based sparse representation," in *Proc. AAAI*, 2011, pp. 313–318.

[30] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1385–1393.
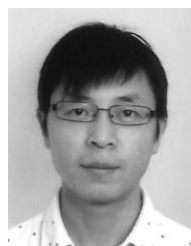
[31] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell. (AAAI)*, Quebec City, QC, Canada, 2014, pp. 2177–2183.

[32] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3864–3872.

[33] S. Moran and V. Lavrenko, "Regularised cross-modal hashing," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 907–910.

[34] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 447–461, Mar. 2016.

[35] M. Yu, L. Shao, X. Zhen, and X. He, "Local feature discriminant projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2015.2497686.

[36] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 679–686.

[37] Y. Wang, X. Lin, L. Wu, W. Zhang, and Q. Zhang, "LBMCH: Learning bridging mapping for cross-modal hashing," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 999–1002.

[38] Y. Pan, T. Yao, H. Li, C.-W. Ngo, and T. Mei, "Semi-supervised hashing with semantic confidence for large scale visual search," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 53–62.

[39] Y. Mu, J. Shen, and S. Yan, "Weakly-supervised hashing in kernel space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3344–3351.

[40] M. Zheng *et al.*, "Graph regularized sparse coding for image representation," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1327–1336, May 2011.

[41] X. Lu, H. Yuan, P. Yan, Y. Yuan, and X. Li, "Geometry constrained sparse coding for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1648–1655.

[42] R. Jiang, H. Qiao, and B. Zhang, "Speeding up graph regularized sparse coding by dual gradient ascent," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 313–317, Mar. 2015.

[43] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 940–948.

[44] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[45] G. Bouchard, D. Yin, and S. Guo, "Convex collective matrix factorization," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, 2013, pp. 144–152.

[46] S.-G. Lee and Q.-P. Vu, "Simultaneous solutions of Sylvester equations and idempotent matrices separating the joint spectrum," *Linear Algebra Its Appl.*, vol. 435, no. 9, pp. 2097–2109, 2011.

[47] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th Int. Conf. Multimedia*, 2010, pp. 251–260.

[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "Nus-wide: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 48.

[49] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2014, pp. 395–404.

Jun Tang is currently a Professor with the School of Electronics and Information Engineering, Anhui University, Hefei, China. His research interests include computer vision, machine learning, and pattern recognition.

Ke Wang is currently pursuing the M.S. degree with the School of Electronics and Information Engineering, Anhui University, Hefei, China. His research interests include computer vision, machine learning, and pattern recognition.

Ling Shao (M'09–SM'10) is currently a Professor with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K. He was a Senior Lecturer (2009-2014) with the Department of Electronic and Electrical Engineering, The University of Sheffield, and a Senior Scientist (2005-2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing, and machine learning. He is a fellow of the British Computer Society and the Institution of Engineering and Technology. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals.