IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 29, 2018, accepted April 27, 2018, date of publication May 2, 2018, date of current version June 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2832141

Supervised Intra- and Inter-Modality Similarity Preserving Hashing for Cross-Modal Retrieval

ZHIKUI CHEN[1,2], (Member, IEEE), FANGMING ZHONG[1], GEYONG MIN[3], YONGLIN LENG[1], AND YIMING YING[4]

[1]School of Software Technology, Dalian University of Technology, Dalian 116620, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian 116620, China
[3]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, U.K.
[4]Department of Mathematics and Statistics, The State University of New York at Albany, Albany, NY 12222, USA

Corresponding author: Zhikui Chen (zkchen@dlut.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2016YFD0800300.

ABSTRACT Cross-modal hashing has drawn considerable interest in multimodal retrieval due to the explosive growth of big data on multimedia. However, the existing methods mainly focus on unified hash codes learning and investigate the local geometric structure in the original space, resulting in low-discriminative power hash code of out-of-sample instances. To address this important problem, this paper is dedicated to investigate the hashing functions learning by considering the modality correlation preserving in the expected low-dimensional common space. A cross-modal hashing method based on supervised collective matrix factorization is proposed by taking intra-modality and inter-modality similarity preserving into account. For more flexible hashing functions, label information is embedded into the hashing functions learning procedure. Specifically, we explore the intra-modality similarity preserving in the expected low-dimensional common space. In addition, a supervised shrinking scheme is used to enhance the local geometric consistency in each modality. The proposed method learns unified hash codes as well as hashing functions for different modalities; the overall objective function, consisting of collective matrix factorization and intra- and inter-modality similarity embedding, is solved using an alternative optimization in an iterative scheme. Extensive experiments on three benchmark data sets demonstrate that the proposed method is more flexible to new coming data and can achieve superior performance to the state-of-the-art supervised cross-modal hashing approaches in most of the cases.

INDEX TERMS Cross-modal retrieval, matrix factorization, similarity preserving hashing, alternative optimization.

I. INTRODUCTION

With the explosive growth of multimedia data pouring into the Internet, cross-modal retrieval has attracted considerable research interests in recent years. It has been an important research topic in many real life applications, such as visual search [1], image captioning [2], and machine translation [3]. Typical example is image search via a text query to return the relevant images sharing the similar semantics. However, the natural heterogeneous gap among different modalities, such as images and texts, makes it difficult to measure the similarity directly. In addition, the retrieval in large-scale datasets becomes quite challenging due to the increasing multimedia data. Most of the existing approaches to addressing these problems project the data from different modalities into a low-dimensional common subspace, in which the similarities can be computed directly. Among these methods, hashing has gained increasing attentions due to its efficiency on retrieval and low storage cost. The main goal of hashing is to transform high-dimensional data into compact binary codes [4]. Thus, the similarities between query code and the retrieval sets can be measured by the Hamming distance, which can be calculated efficiently via fast bit-wise XOR operation in the expected Hamming space [5]. The retrieval results are selected from the instances ranked in an ascending order in terms of Hamming distances. Hence, in order to boost the cross-modal retrieval, it is essential to learn the hashing function that can well embed the correlation across different modalities and the relationship within each modality.

Recently, significant efforts have been shifted to hashing function learning for cross-modal retrieval. Linear

Cross-Modal Hashing (LCMH) [6] was proposed based on a clustering representation to preserve the inter-similarity among different modalities and intra-similarity in each modality. Inter-media hashing (IMH) algorithm was proposed in [7] to facilitate large-scale cross-media retrieval. IMH explored both the intra-media and inter-media consistencies to achieve effective binary codes. It learns a set of hashing functions for each bite of the hash code in each individual modality. Inter-media consistency is formulated by the shared similar semantics, and intra-media consistency is formulated by the local structure information i.e. affinity relationship within each individual modality. Cross-view hashing (CVH) [5] formulated the problem of learning hashing functions as a generalized eigenvalue problem. Supervised Multimodal Hashing (SMH) [8] was proposed to seamlessly integrate semantic labels into the hashing leaning procedure for large-scale data modeling. Most recently, Collective Matrix Factorization Hashing (CMFH) [9] uses matrix factorization to learn the latent concepts from each modality, which has achieved an impressive result on cross-modal retrieval. Nonetheless, CMFH does not preserve the local structure information within each individual modality i.e. intra-modality similarity, nor takes the inter-modality correlations into consideration. Inspired by CMFH, several extensions based on matrix factorization have been proposed to formulate the supervised label information, such as supervised matrix factorization (SMFH) [10], [11], cluster-based joint matrix factorization (C-JMFH) [12], Supervised CMFH (SCMFH) [13], etc. In particular, SMFH [10] integrates the graph regularization into the collective non-negative matric factorization. Furthermore, label information is used to refine the graph regularizer. Different from CMFH, SMFH uses multiplicative iteration and a subsample method as an efficient optimization algorithm. As the supervised label information is taken in to account, SMFH learns more discriminative hash codes and achieves the superior retrieval performance. According to the availability of label information, most of these approaches can be briefly categorized into two groups i.e., unsupervised methods [9], [12], [14], [15], and supervised methods [5], [10], [15]. Generally, the label information is beneficial to enhance the correlation among different modalities for unified hash codes in the Hamming space. For instance, image and text instances with the same label information should share similar hash codes in the Hamming space. By so doing, discrimination and relevance of the learned hash codes can be further strengthened. Hence, supervised cross-modal hashing methods can usually yield the better performance. However, it still remains unclear how to formulate the intra-modality and inter-modality correlations using the label information in the supervised setting.

One fundamental limitation of the most existing approaches is that they only formulated the supervised label information for common semantics learning ignoring the hashing function learning. In addition, most existing methods study the local geometric structure in the original feature space. To the best of our knowledge, there are few cross-modal hashing methods which investigate local structure preservation in the expected low-dimensional common space. To fill this gap, we aim to enhance the discriminative power of hash codes in each individual modality in the learned Hamming space for better retrieval performance. A natural way to define local consistency in a hashing function learning problem is to directly define consistency i.e. neighborhood relationship based on the hash codes in the expected low-dimensional Hamming space.

To address these challenges, we propose a novel intra-modality and inter-modality similarity preserving hashing approach based on supervised matrix factorization and graph regularization termed IISPH. It is devoted to learning more flexible hashing functions that can preserve the inter-modality similarity across different modalities, as well as the local geometric structure in each modality in the expected low-dimensional common space. We first utilize collective matrix factorization to learn latent concept from each modality. For out-of-sample instances, we learn a linear projection as the hashing function for each modality respectively. Then we take intra-modality and inter-modality similarity preservation into consideration. Different from SMFH [10], we investigate intra-modality similarity in the expected low-dimensional common space. Furthermore, we incorporate the hashing functions learning to the similarities formulation, which eventually boils down to two graph regularization terms. In addition, supervised label information is used to refine the local neighborhood structure in the expected low-dimensional common space. The objective function is solved by alternative optimization in an iterative manner. We conjecture that all these would facilitate preserving the global and local structure of original feature space during hashing functions learning and improving the cross-modal retrieval performance.

The major contributions of this paper are summarized as follows.

- We investigate both intra-modality and inter-modality similarity preservation in the learning of compact hash codes. Since similar instances should share the same hash code and supervised label information can facilitate the discriminative hash codes learning, we enable the similarity preservation based supervised collective matrix factorization and linear transformation. As a result, the discriminative power of learned hash codes can be strengthened.
- Different from the existing cross-modal hashing methods, we embed the supervised information into the formulation of intra-modality and inter-modality similarity preservation. Due to the exploitation of new coming multimedia data, we combine the hashing functions learning and similarity preservation with supervised label information to learn more flexible hashing functions for out-of-sample data. It is essential for the performance improvement and time complexity decreasing of cross-modal retrieval.

- Furthermore, preserving intra-modality similarity is investigated by means of local geometric structure under the expected low-dimensional common space. Aiming at preserving the local consistency, we formulate the intra-modality similarity preservation by replacing the original space data with the data under the expected space. To the best of our knowledge, this is the first time to extend the intra-modality similarity preservation to the expected space for cross-modal hashing. This is the main difference between our method and the existing methods. Extensive experimental results demonstrate the superior effectiveness of the proposed IISPH.

The rest of this paper is organized as follows. Section II reviews the existing work on cross-modal hashing. We present the detailed information of the proposed method in Section III. In Section IV, extensive experimental details and results are described in comparison with state-of-the-art methods on three benchmark datasets. Finally, this work is concluded in Section V.

II. RELATED WORK

In this section, previous efforts on cross-modal retrieval will be reviewed and analyzed based on the utilization of label information and the consideration of intra-modality and inter-modality similarity preservation.

A. COMMON SUBSPACE LEARNING

In order to perform cross-modal retrieval, common subspace for bridging the heterogeneous gap across different modalities is the most widely used approach in the last decade, where the similarity can be directly measured [16], [17]. Many subspace learning based methods have been proposed to learn a latent common space for cross-modal retrieval. In regard to the utilization of label information, the previous subspace learning approaches can be broadly classified into two groups, i.e., unsupervised and supervised subspace learning, respectively. Canonical Correlation Analysis (CCA) [18], [19] is one of the most popular cross-modal subspace learning methods. CCA transforms the original feature to a low-dimensional latent subspace with the correlation of paired samples being maximized. Thus a pair of instances from different modalities will have the similar representations in the subspace, which is beneficial to cross-modal retrieval. The limitation of CCA is that, the correlations among different pairs and items in different pairs are not considered in the subspace learning. Additionally, CCA is an unsupervised method, without considering the label information. Different from CCA, Kang *et al.* [20] proposed a supervised consistent feature representation learning method with the capability of dealing with unpaired training samples. He *et al.* [21] proposed a cross-modal matching methods based on compound ℓ_{21} regularization to reduce the semantic gap and outliers under the condition of pairwise constraints. Both supervised and unsupervised schemes were investigated in [21]. In [22], a joint graph regularized multimodal subspace learning approach was proposed to better explore the cross-modal correlation and the local manifold structure. In [23], ℓ_{21}-norms and graph regularization are coupled with a linear regression to learn projection matrices for mapping different modal data into the common space. With these coupled items, the approach proposed in [23] can preserve the relationships of instances from different modalities, also the relationships among instances in each individual modality.

B. CROSS-MODAL HASHING

Due to the efficiency of retrieval and low storage cost, hashing methods are widely investigated. Hashing was incorporated into the subspace learning methods. Thus, the cross-modal retrieval is formulated as the problem of hashing function learning. The feature representations in low-dimensional common subspace are transformed to more compact hash codes under a Hamming space. Similar to the subspace learning methods, cross-modal hashing methods can be briefly classified into two groups too, i.e., taking inter-modality similarity preservation into account only, and taking both inter-modality and intra-modality similarities preservation into consideration.

1) UNSUPERVISED INTER-MODALITY SIMILARITY PRESERVATION

Inter-modality similarity is considered in many cross-modal hashing methods for better learning the common semantic relationship. Among them, a few methods are unsupervised [9], [14], and they measure the inter-modality relationship by training the paired samples which describe the same object. Ding *et al.* [9] firstly proposed collective matrix factorization to obtain the latent concepts in different modalities. Although CMFH does not consider the label information, it has achieved an impressive performance on cross-modal retrieval that demonstrated the power of matrix factorization in latent structure learning. Zhou *et al.* [14] proposed Latent Semantic Sparse Hashing (LSSH), which learns the semantic concepts of images and text by sparse coding and matrix factorization, respectively. The learned latent semantic features are then mapped to a joint abstraction space where the unified hash codes are generated via quantization. However, the weakness of these methods is that the supervised information has not been fully used.

2) SUPERVISED INTER-MODALITY SIMILARITY PRESERVATION

Others considered the label information to enhance the common semantic relationship. Cross-view hashing (CVH) [5] is extended from spectral hashing [24], which formulated the hashing functions learning as a generalized eigenvalue problem. Zhang and Li [8] proposed a semantic correlation maximization (SCM) to embed label information into the hashing learning procedure seamlessly. SCM solves the hashing functions learning by learning the orthogonal projections. Ding *et al.* [13] extended CMFH to the supervised scenario where it fully exploits the label information of data, termed as SCMFH. SCMFH formulated the unified hash code learning in the shared latent semantic space as a joint optimization with classification problem. With the label information taken in to account, SMFH [10] constructs a graph regularization to investigate the correlations across different modalities. Different from CMFH, it uses multiplicative iteration and a subsample method for efficient optimization. Thanks to the contribution of matrix factorization and supervised label information, SMFH has obtained the superior retrieval performance.

By integrating the label information, users can embed the must-link or must-not-link constraints into the hashing functions learning. For instance, the samples from different modalities with the same label information should share similar representation in the Hamming space. Thus, the learned hashing functions will be more powerful to extract the intrinsic semantics across different modalities, as well as to improve the cross-modal retrieval performance. However, a limitation of the above methods is that, the relationships among samples in the same modality are not considered, i.e. intra-modality similarity.

3) INTRA-MODALITY SIMILARITY PRESERVATION

In order to preserve the intra-modality relationship in each individual modality, most existing methods formulated it as a graph regularization. In [7], IMH explores the intra-media consistency by the affinity relationship in each modality. Rafailidis and Crestani [12] proposed cluster-based joint matrix factorization hashing (C-JMFH) to learn cross-modal cluster representations for instances, which are incorporated into the joint matrix factorization later to measure the inter-modality and intra-modality similarities. In [15], Wang *et al.* proposed leaning bridging mapping for cross-modal hashing (LBMCH) to explore the semantic correspondence of distinct Hamming spaces which can characterize the discriminative local structure for each modality. IMH, C-JMFH, and LBMCH are unsupervised methods, which do not make full use of label information. There are also many supervised cross-modal hashing approaches that take the intra-modality similarity into consideration. In [25], Co-Regularized Hashing (CRH) based on a boosted co-regularization framework is proposed. CRH learns hashing function for each bit of the hash codes from each modality. Wu *et al.* [26] proposed Sparse Multimodal Hashing (SM^{2}H) to model both intra-modality and inter-modality similarities as a hypergraph. In [11], Tang *et al.* formulated the label consistency across different modalities and the local geometric consistency in each modality as a mixed graph regularization term. However, these methods only modeled the intra-modality similarity in the original space.

C. SUMMARY

By reviewing the previous work, we can observe that, the previous cross-modal hashing methods mostly concentrated on leaning the unified hash codes from the given training data, and the supervised label information has only been embedded into the common semantics learning rather than the hashing functions learning. Motivated by the promising results delivered by matrix factorization in cross-modal retrieval [9]–[11], we make further efforts to investigate the hashing function learning incorporated with supervised label information, and explore the local geometric structure in the expected Hamming space in a natural way.

III. PROPOSED IISPH

In this section, we introduce our method called IISPH. The framework of IISPH is shown in Fig. 1. Different from the existing methods, we embed the label information to the hashing functions learning, rather than the learning of unified hash codes from training data. Additionally, the intra-modality similarity in this paper is considered in the expected low-dimensional common space.

As shown in Fig. 1, the supervised label information is utilized to learn hashing functions from training data. The goal is to project the original data from different modalities into a unified Hamming space, where the similarity can be calculated directly. The framework consists of unified hash code leaning and hashing functions learning, as well as the intra-modality and inter-modality similarity preservation.

A. PROBLEM FORMULATION

Several important notations used in this paper are illustrated in Table 1. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as $\|\mathbf{A}\|_F = \sqrt{tr(\mathbf{A}^T\mathbf{A})} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$. Given cross-modal data $\mathbf{X}^{(1)} = \{x_1^1, \cdots, x_n^1\}$ and $\mathbf{X}^{(2)} = \{x_1^2, \cdots, x_n^2\}$, such as images and the associated text. We have n samples from each modality, that $\mathbf{X}^{(1)} \in \mathbb{R}^{d_1\times n}, \mathbf{X}^{(2)} \in \mathbb{R}^{d_2\times n}$, where d_1 represents the dimension of image feature, and d_2 denotes the dimension of text descriptor (usually $d_1 \neq d_2$). Without loss of generality, we assume that the cross-modal data are zero-centered, i.e., $\sum_{i=1}^n x_i^1 = \mathbf{0}$ and $\sum_{i=1}^n x_i^2 = \mathbf{0}$.

The goal of cross-modal hashing is to learn a hashing function for each modality data, which can transform data from the original space to a Hamming space. The hashing function can be defined as:

$$H^t : \mathbb{R}^{d_t} \mapsto \{-1, 1\}^k, \quad t = 1, 2, \tag{1}$$

where k is the length of hash code. Here, we use $\{-1, 1\}$ to represent hash codes $\mathbf{Y}$, which can be easily transformed to binary codes via mean thresholding stated as follows:

$$\mathbf{H}^t = \frac{1}{2}(\mathbf{1} + \mathbf{Y}^t), \quad t = 1, 2, \tag{2}$$

where $\mathbf{Y}^t = sign(\mathbf{V}^t)$, $\mathbf{V}$ is the real-valued representation of low-dimensional space. $sign$ is the sign function $sign(u) = 1$ if $u > 0$ and -1 otherwise for all $u \in \mathbb{R}$. For convenience, we define the hash codes as $\{-1, 1\}^k$ in the rest of this paper.

In order to preserve the intra-modality and inter-modality similarities in the expected Hamming space, we incorporate supervised label information to hashing function learning. Thus we also have the label information of given cross-modal data. Let $\mathbf{C}$denote the label matrix $\mathbf{C} \in \mathbb{R}^{c\times n}$, where c is the

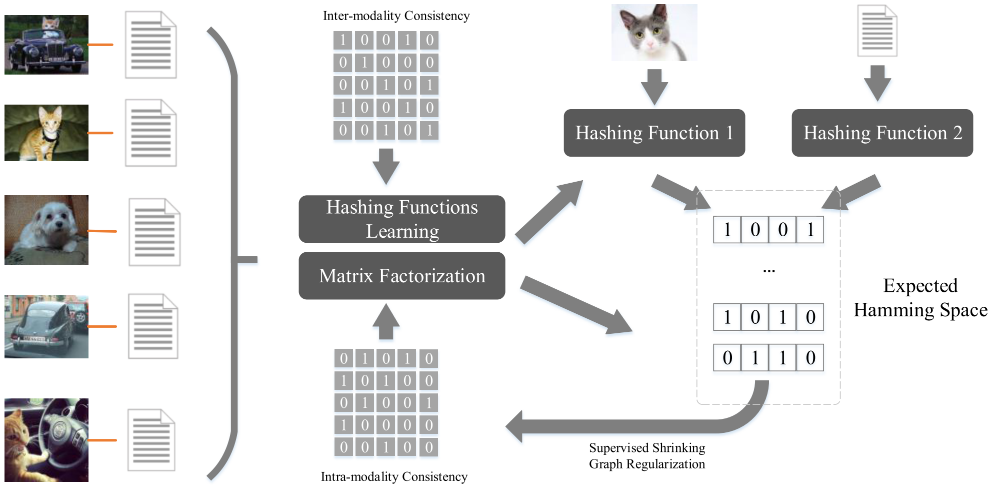


FIGURE 1. The framework of the proposed method.

TABLE 1. Important notations used in this paper.

Symbol	Definition
$\mathbf{X}$	Data matrix
$\mathbf{V}$	Common semantic representation
$\mathbf{Y}$	Unified hash codes
$\mathbf{P}$	Projection matrix
$\mathbf{U}$	Latent factor
$\mathbf{A}$	Inter-modality similarity
$\mathbf{S}$	Intra-modality similarity
$\mathbf{L}$	Laplacian matrix
$\mathbf{C}$	Label matrix
n	The number of samples
c	The number of classes
k	The length of hash codes
d_1, d_2	Dimensionality of cross-modal data

total number of classes. Since the cross-modal data describe the same objects, they share the same label information. The i-th column of $\mathbf{C}$ is $\mathbf{C}_i \in \{0, 1\}^c$, which is the label of image $x_i^{(1)}$ and text $x_i^{(2)}$. We assume that the images and text belong to at least one of the c classes. If the i-th item belongs to the j-th class, $\mathbf{C}(j, i) = 1$, otherwise $\mathbf{C}(j, i) = 0$.

B. COLLECTIVE MATRIX FACTORIZATION

Matrix factorization was first used in [9] to learn the latent semantic concept from the original feature space for cross-modal hashing. Given a matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, it can be decomposed into two factors by matrix factorization as bellow:

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}, \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{k\times n}$ can be considered as the k-dimensional latent space representation of $\mathbf{X}$, and $\mathbf{U} \in \mathbb{R}^{d\times k}$ denotes the low-dimensional representation of features. Generally, in the low dimensional latent space, the compact representation is more effective to measure the similarity. Suppose we have two modalities $\mathbf{X}^{(1)} \in \mathbb{R}^{d_1\times n}, \mathbf{X}^{(2)} \in \mathbb{R}^{d_2\times n}$, and they describe the same object in different views. It is reasonable to conjecture that their latent representation should share the same semantics. Thus, it is expected to extract the common semantics by collective matrix factorization under the constraints that they share the same latent space stated as follows:

$$\mathbf{X}^{(1)} \approx \mathbf{U}_1\mathbf{V}, \tag{4}$$

$$\mathbf{X}^{(2)} \approx \mathbf{U}_2\mathbf{V}, \tag{5}$$

where $\mathbf{U}_1 \in \mathbb{R}^{d_1\times k}$, $\mathbf{U}_2 \in \mathbb{R}^{d_2\times k}$, and $\mathbf{V} \in \mathbb{R}^{k\times n}$ represents the common semantics. If we use squared Frobenius norm as the loss function of $\mathbf{X}$ and $\mathbf{UV}$, the common semantics learning of cross-modal data can be formulated as:

$$\begin{aligned} &O_1(\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}) \\ &\quad = \alpha\|\mathbf{X}^{(1)} - \mathbf{U}_1\mathbf{V}\|_F^2 + (1-\alpha)\|\mathbf{X}^{(2)} - \mathbf{U}_2\mathbf{V}\|_F^2, \end{aligned} \tag{6}$$

where α denotes the balance parameter that weights the modality importance, and $\|\bullet\|_F^2$ represents the squared Frobenius norm.

The learned common semantics in (6) are only appropriate to the training data. For the new coming instances, it is computationally expensive to retrain the whole data set. Hence, hashing functions mapping data from the original feature space to the common latent space should be learned for out-of-sample instances. Here, we follow the idea of learning two linear projections as the hashing functions [11]. Then we have:

$$\mathbf{V}_1 = \mathbf{P}_1\mathbf{X}^{(1)}, \tag{7}$$

$$\mathbf{V}_2 = \mathbf{P}_2\mathbf{X}^{(2)}, \tag{8}$$

where $\mathbf{P}_1 \in \mathbb{R}^{k\times d_1}$, and $\mathbf{P}_2 \in \mathbb{R}^{k\times d_2}$. As mentioned above, two modality data shave the same semantics, thus we have $\mathbf{V}_1 = \mathbf{V}_2$. Hence, the hashing functions learning can be formulated as:

$$O_2(\mathbf{P}_1, \mathbf{P}_2) = \|\mathbf{V} - \mathbf{P}_1\mathbf{X}^{(1)}\|_F^2 + \|\mathbf{V} - \mathbf{P}_2\mathbf{X}^{(2)}\|_F^2. \tag{9}$$

By minimizing (9), the two projection matrices that map cross-modal data onto a common semantic space will be learned.

C. INTRA- AND INTER-MODALITY SIMILARITY EMBEDDING

In order to improve the cross-modal retrieval performance, many cross-modal hashing methods have taken the intra-modality similarity preservation into consideration. Similar to [11] and [23], we use graph Laplacian regularization for preserving the consistency. These methods define the local consistency in the original feature space, whereas the goal of cross-modal hashing is to learn a compact Hamming space, thus the original feature space is not the best for defining local consistency. Our method differs from them in that IISPH explores the local geometric structure in the expected Hamming space. Thus, the obtained hash codes preserve the local consistency in a natural way.

We use Laplacian Eigenmaps (LE) [27] to formulate the similarity preservation based on manipulations on an undirected weight graph which indicates the neighboring relationships of pairwise data. Hence, the objective function of intra-modality similarity preservation in the expected Hamming space can be stated as follows:

$$\min_{\mathbf{P}_1,\mathbf{P}_2} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{2} \|\mathbf{P}_t\mathbf{X}_i^{(t)} - \mathbf{P}_t\mathbf{X}_j^{(t)}\|^2\mathbf{S}_{ij}^{(t)}, \tag{10}$$

where $\mathbf{S}_{ij}^{(t)}$, $t = 1, 2$, is the affinity matrix in the t-th modality under the expected low-dimensional space which is defined as below:

$$\mathbf{S}_{ij}^{(t)} = \begin{cases} \exp(\frac{-\|x_i^{(t)} - x_j^{(t)}\|^2}{2\sigma^2}), \\ \quad \text{if } x_i^{(t)} \in \mathbb{N}_k(x_j^{(t)}) \text{ or } x_j^{(t)} \in \mathbb{N}_k(x_i^{(t)}) \\ 0, \quad \text{otherwise} \end{cases} \tag{11}$$

where $\|x_i^{(t)} - x_j^{(t)}\|^2$ is the Euclidean distance between samples $x_i^{(t)}$ and $x_j^{(t)}$, σ is the median value of the distances matrix in each modality, and $\mathbb{N}_k(x_j^{(t)})$ is the k-nearest neighbors of $x_j^{(t)}$.

From the existing work [11], [15], [23], it can be observed that the existing methods do not take the supervised label information into account. Since the k-nearest neighbors may have different class labels, we incorporate the label information [28] to shrink the distance matrix by:

$$\mathbf{Dist}_{ij}^{(t)} = \begin{cases} \exp(-\frac{\mathbf{Dist}_{ij}^{(t)}}{\rho\xi}) \times \mathbf{Dist}_{ij}^{(t)}, & \text{if } (\mathbf{C}_i^{(t)})^T\mathbf{C}_j^{(t)} > 1 \\ \mathbf{Dist}_{ij}^{(t)}, & \text{otherwise} \end{cases} \tag{12}$$

where $(\mathbf{C}_i^{(t)})^T\mathbf{C}_j^{(t)} > 1$ denotes that $x_i^{(t)}$ and $x_j^{(t)}$ have at least one common class label, ρ and ξ are applied to prevent the pairwise distance from decreasing too fast. ξ takes the average Euclidean distance of $\mathbf{Dist}^{(t)}$ and $0 < \rho < 1$.

Since we consider the local geometric information in the expected Hamming space, it can be seen that this scheme poses a chicken-and-egg problem [29]. The projection matrix $\mathbf{P}$ needs to be computed based on the affinity matrix $\mathbf{S}$, but $\mathbf{S}$ is calculated under the projection space of $\mathbf{P}$. To address this problem, we adopt a greedy approach that given $\mathbf{P}^{(t-1)}$ in the $(t-1)$ iteration, we can use $\mathbf{P}^{(t-1)}$ as an approximation of $\mathbf{P}^{(t)}$ to calculate $\mathbf{S}$ in the t-th iteration.

Hence, the objective of intra-modality similarity preservation can be stated as:

$$O_3(\mathbf{P}_1, \mathbf{P}_2) = \frac{1}{2}\sum_{t=1}^{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{P}_t\mathbf{X}_i^{(t)} - \mathbf{P}_t\mathbf{X}_j^{(t)}\|^2\mathbf{S}_{ij}^{(t)}. \tag{13}$$

Through algebraic calculation, the objective function in (13) can be reformulated as:

$$\begin{aligned} O_3(\mathbf{P}_1, \mathbf{P}_2) &= \mathrm{tr}[\mathbf{P}_1\mathbf{X}^{(1)}\mathbf{L}_1(\mathbf{X}^{(1)})^T\mathbf{P}_1^T + \mathbf{P}_2\mathbf{X}^{(2)}\mathbf{L}_2(\mathbf{X}^{(2)})^T\mathbf{P}_2^T] \\ &= \mathrm{tr}(\mathbf{Q}\tilde{\mathbf{L}}\mathbf{Q}^T), \end{aligned} \tag{14}$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are the Laplacian matrix of $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ respectively. $\mathbf{L}_1 = \mathbf{D}_1 - \mathbf{S}^{(1)}$, where $\mathbf{D}_1 \in \mathbb{R}^{n\times n}$ is a diagonal matrix. The diagonal entries of $\mathbf{D}_1$ are the column sum of $\mathbf{S}^{(1)}$, i.e. $\mathbf{D}_1(i, i) = \sum_j \mathbf{S}_{ij}^{(1)}$. Similarly, $\mathbf{L}_2 = \mathbf{D}_2 - \mathbf{S}^{(2)}$, where $\mathbf{D}_2 \in \mathbb{R}^{n\times n}$, and $\mathbf{D}_2(i, i) = \sum_j \mathbf{S}_{ij}^{(2)}$. Here, $\mathbf{Q} = [\,\mathbf{P}_1\mathbf{X}^{(1)}\ \mathbf{P}_2\mathbf{X}^{(2)}\,]$, and $\tilde{\mathbf{L}} = \begin{bmatrix} \mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2 \end{bmatrix}$.

We then consider the inter-modality similarity preservation. In the previous work, this relationship was modeled on the common semantics $\mathbf{V}$. However, we attempt to learn more powerful hashing functions for the large-scale out-of-sample instances by incorporating the label information into the hashing functions learning procedure. Hence, the objective function of inter-modality similarity preservation combined with hashing functions learning can be defined as:

$$\min_{\mathbf{P}_1,\mathbf{P}_2} \sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{P}_1\mathbf{X}_i^{(1)} - \mathbf{P}_2\mathbf{X}_j^{(2)}\|^2\mathbf{A}_{ij}, \tag{15}$$

where $\mathbf{A}$ is the affinity matrix across two modalities $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. $\mathbf{A}$ is computed based on the label information as follows:

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if } \mathbf{X}_i^{(1)} \text{ and } \mathbf{X}_j^{(2)} \text{ have the same label} \\ 0. & \text{otherwise} \end{cases} \tag{16}$$

Through algebraic calculation, the objective function in (15) can be reformulated as:

$$\begin{aligned} &O_4(\mathbf{P}_1, \mathbf{P}_2) \\ &\quad = \mathrm{tr}[\mathbf{P}_1\mathbf{X}^{(1)}\mathbf{D}_{12}(\mathbf{X}^{(1)})^T\mathbf{P}_1^T + \mathbf{P}_2\mathbf{X}^{(2)}\mathbf{D}_{21}(\mathbf{X}^{(2)})^T\mathbf{P}_2^T] \\ &\qquad - \mathrm{tr}[\mathbf{P}_1\mathbf{X}^{(1)}\mathbf{A}_{12}(\mathbf{X}^{(2)})^T\mathbf{P}_2^T + \mathbf{P}_2\mathbf{X}^{(2)}\mathbf{A}_{21}(\mathbf{X}^{(1)})^T\mathbf{P}_1^T] \\ &\quad = \mathrm{tr}(\mathbf{Q}\mathbf{L}'\mathbf{Q}^T), \end{aligned} \tag{17}$$

where $\mathbf{D}_{12}$ is the diagonal matrix whose entries are $\mathbf{D}_{12}(i, i) = \sum_j \mathbf{A}_{12}(i, j)$, $\mathbf{D}_{12} = \mathbf{D}_{21}^T$, and $\mathbf{A}_{12} = \mathbf{A}_{21}^T$ are the affinity matrices defined as (16). $\mathbf{Q} = [\,\mathbf{P}_1\mathbf{X}^{(1)}\ \mathbf{P}_2\mathbf{X}^{(2)}\,]$, $\mathbf{L}' = \mathbf{D}' - \mathbf{A}'$

is the Laplacian matrix, $\mathbf{D}'$ is a diagonal matrix whose entries are $\mathbf{D}'_{ii} = \sum_j \mathbf{A}'_{ij}$, and $\mathbf{A}'$ is defined as follows:

$$\mathbf{A}' = \begin{bmatrix} \mathbf{0} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{0} \end{bmatrix}. \tag{18}$$

So far we have formulated the intra-modality and inter-modality similarity preservation incorporated with supervised label information. Additionally, hashing functions learning is combined to the similarity preserving procedure.

D. OVERALL OBJECTIVE FUNCTION

Consisting of collective matrix factorization, hashing functions learning, and intra-modality and inter-modality similarity preservation, the overall objective function of our proposed IISPH can be defined as follows:

$$\begin{aligned} \min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{P}_1,\mathbf{P}_2,\mathbf{V}} \; & O(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{V}) \\ = & \; O_1 + \beta O_2 + \lambda O_3 + \mu O_4 + \gamma R(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{V}) \\ = & \; \alpha \|\mathbf{X}^{(1)} - \mathbf{U}_1\mathbf{V}\|_F^2 + (1-\alpha)\|\mathbf{X}^{(2)} - \mathbf{U}_2\mathbf{V}\|_F^2 \\ & + \beta(\|\mathbf{V} - \mathbf{P}_1\mathbf{X}^{(1)}\|_F^2 + \|\mathbf{V} - \mathbf{P}_2\mathbf{X}^{(2)}\|_F^2) \\ & + \lambda \mathrm{tr}(\mathbf{Q}\tilde{\mathbf{L}}\mathbf{Q}^T) + \mu \mathrm{tr}(\mathbf{Q}\mathbf{L}'\mathbf{Q}^T) \\ & + \gamma R(\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{V}), \end{aligned} \tag{19}$$

where β, λ, μ and γ are trade-off parameters of the corresponding terms. $R(\cdot) = \|\cdot\|_F^2$ denotes the regularization term to avoid overfitting.

Algorithm 1 Intra- and Inter-Modality Similarity Preserving Hashing

```
Input:
    Cross-modal data $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, parameters $\beta$, $\lambda$, $\mu$, $\gamma$,
    the number of of the $k$-nearest neighbors, label matrix $\mathbf{C}$,
    and the length of hash codes $k$.
Output:
    Unified hash codes $\mathbf{Y}$, projection matrix $\mathbf{P}_1, \mathbf{P}_2$.
1. Center $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ by means, and construct the Laplacian matrix $\mathbf{L}'$.
2. Initialize $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}$ by random matrices respectively,
    $\mathbf{P}_1 = \mathbf{I}_{k\times d_1}, \mathbf{P}_2 = \mathbf{I}_{k\times d_2}$.
3. Compute $\tilde{\mathbf{L}}$ by (11) and (12) based on $\mathbf{P}_1, \mathbf{P}_2$.
4. repeat
5.    Fix $\mathbf{P}_1, \mathbf{P}_2, \mathbf{V}$, update $\mathbf{U}_1, \mathbf{U}_2$ by (20) and (21).
6.    Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2$, update $\mathbf{V}$ by (22).
7.    Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}$, update $\mathbf{P}_1, \mathbf{P}_2$ by (23) and (24).
8.    Update $\tilde{\mathbf{L}}$ by (11) and (12) based on $\mathbf{P}_1, \mathbf{P}_2$.
9. until convergence.
10. $\mathbf{Y} = sign(\mathbf{V})$.
```

E. OPTIMIZATION

Since the optimization problem in (19) is non-convex with five matrix variables $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{V}$, it is intractable to be directly minimized. Fortunately, it is convex with respect to any of the five variables in the case that the others are fixed. Therefore, we employ an alternative optimization in an iterative manner to address the optimization problem until convergence. The detailed optimization steps are listed as follows:

Step 1. Fix $\mathbf{P}_1, \mathbf{P}_2, \mathbf{V}$, then update $\mathbf{U}_1, \mathbf{U}_2$. Let $\frac{\partial O}{\partial \mathbf{U}_1} = 0$ and $\frac{\partial O}{\partial \mathbf{U}_2} = 0$, we can have:

$$\mathbf{U}_1 = \mathbf{X}^{(1)}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T + \frac{\gamma}{\alpha}\mathbf{I})^{-1}, \tag{20}$$

$$\mathbf{U}_2 = \mathbf{X}^{(2)}\mathbf{V}^T(\mathbf{V}\mathbf{V}^T + \frac{\gamma}{1-\alpha}\mathbf{I})^{-1}. \tag{21}$$

Step 2. Fix $\mathbf{P}_1, \mathbf{P}_2, \mathbf{U}_1, \mathbf{U}_2$, then update $\mathbf{V}$. Let $\frac{\partial O}{\partial \mathbf{V}} = 0$, we can have:

$$\begin{aligned} \mathbf{V} = & [\alpha \mathbf{U}_1^T\mathbf{U}_1 + (1-\alpha)\mathbf{U}_2^T\mathbf{U}_2 + (2\beta + \gamma)\mathbf{I}]^{-1} \\ & \times [\alpha \mathbf{U}_1^T\mathbf{X}^{(1)} + (1-\alpha)\mathbf{U}_2^T\mathbf{X}^{(2)} + \beta\mathbf{P}_1\mathbf{X}^{(1)} + \beta\mathbf{P}_2\mathbf{X}^{(2)}]. \end{aligned} \tag{22}$$

Step 3. Fix $\mathbf{U}_1, \mathbf{U}_2, \mathbf{V}$, update $\mathbf{P}_1, \mathbf{P}_2$. Let $\frac{\partial O}{\partial \mathbf{P}_1} = 0$ and $\frac{\partial O}{\partial \mathbf{P}_2} = 0$, we can have:

$$\begin{aligned} \mathbf{P}_1 = & [\beta\mathbf{V}(\mathbf{X}^{(2)})^T + \mu\mathbf{P}_2\mathbf{X}^{(2)}\mathbf{A}_{21}(\mathbf{X}^{(1)})^T] \\ & \times [\beta\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T + \mu\mathbf{X}^{(1)}\mathbf{D}_{12}(\mathbf{X}^{(1)})^T \\ & + \lambda\mathbf{X}^{(1)}\mathbf{L}_1(\mathbf{X}^{(1)})^T + \gamma\mathbf{I}]^{-1}. \end{aligned} \tag{23}$$

$$\begin{aligned} \mathbf{P}_2 = & [\beta\mathbf{V}(\mathbf{X}^{(2)})^T + \mu\mathbf{P}_1\mathbf{X}^{(1)}\mathbf{A}_{12}(\mathbf{X}^{(2)})^T] \\ & \times [\beta\mathbf{X}^{(2)}(\mathbf{X}^{(2)})^T + \mu\mathbf{X}^{(2)}\mathbf{D}_{21}(\mathbf{X}^{(2)})^T \\ & + \lambda\mathbf{X}^{(2)}\mathbf{L}_2(\mathbf{X}^{(2)})^T + \gamma\mathbf{I}]^{-1}. \end{aligned} \tag{24}$$

The overall procedure of our IISPH is summarized in Algorithm 1.

F. COMPLEXITY ANALYSIS

This section analyses the complexity of the proposed method. In the intra-modality and inter-modality similarity preservation, constructing Laplacian matrix of intra-modality similarity takes $O((d+k)n^2)$, where $d = \max\{d_1, d_2\}$ and constructing Laplacian matrix of inter-modality similarity takes $O(n^2)$. In addition, in the iteration of updating $\mathbf{U}_1, \mathbf{U}_2, \mathbf{P}_1, \mathbf{P}_2, \mathbf{V}$, two types of inverse computation which costs $O(k^3)$ and $O(d^3)$ respectively. Hence, the time complexity for training IISPH is $O((d^3 + k^3 + nkd + k^2n + dk^2 + (d+k)n^2)T + (d+k+1)n^2)$. Due to the value of $d, k \ll n$, the overall complexity is approximately $O(n^2(d+k)T)$. Furthermore, for online query of an out-of-sample instance, the time complexity scales $O(dk)$, which is dramatically efficient.

IV. EXPERIMENTS AND RESULTS ANALYSIS

In this section, the detailed information about experiments performed to validate the effectiveness of our proposed IISPH will be presented. We conduct the experiments on three representative cross-modal datasets consisting of images and text. In order to evaluate the performance of cross-modal retrieval, we design two cross-modal retrieval tasks, i.e. image to text and text to image. Image to text (**Task 1**) uses image as query to search relevant text, and text to image (**Task 2**) utilizes text as query to search relevant images. In cross-modal retrieval tasks, an image and a text are considered to be relevant when they have the same the semantic label.

A. EXPERIMENTAL SETTINGS

1) DATASETS

Wiki [3] dataset was crawled from the Wikipedia's features articles, which consists of 2866 documents. These documents are image-text pairs which can be grouped into 10 semantic categories. The images are described in 128-dimensional bag-of-visual words SIFT feature vectors, whereas text are represented by 10-dimensional topic vectors generated by the latent Dirichlet allocation (LDA) [30] model. 2173 image-text pairs are randomly selected for training and also as the retrieval set. The remained 693 pairs are used as query set for testing.

Pascal VOC dataset contains 5011 training and 4952 testing image-tag pairs, which can be classified into 20 categories. Since several image-tag pairs are multi-labeled, we select the pairs with only one label as the way in [23] for convenience, resulting in 2808 training and 2841 testing image-tag pairs. The image modality is represented by 512-dimensional GIST features [31], and the representations of text modality are 399-dimensional word frequency features.

MIR Flickr[32] consists of 25000 images collected from Flickr associated with tags. These images belong to at least one of the 24 semantic classes. We select experimental data in the same way reported in [33], which leads to a dataset with 16738 instances. We randomly select 5% (836) for testing, and 5000 instances for training. Here, the images are represented by 150-dimensional edge histogram (EH), and associated tags are described by binary tagging vectors. For convenience, PCA is employed to reduce the dimensionality of text features, resulting in a 500-dimensional feature representation.

2) BASELINES

In order to evaluate the effectiveness of our IISPH, we employ Cross-View Hashing (CVH) [5], Collective Matrix Factorization Hashing (CMFH) [9], Supervised Multimodal Hashing (SMH) [8], and Supervised Matrix Factorization Hashing (SMFH) [10] as baselines for comparison with IISPH. Among these methods, CMFH is unsupervised, CVH, SMH, and SMFH are supervised. Both SMFH and our IISPH are the supervised extension of CMFH, whereas our IISPH models the label information on the hashing function learning and investigates local geometric structure in the expected space. In the experiments, the parameters' setting of all these competitors are based on these presented in their papers.

3) EVALUATION PROTOCOLS

In this paper, we use mean Average Precision (mAP) as the metric to evaluate the performance of cross-modal retrieval, which can be defined as below:

$$mAP = \frac{1}{N}\sum\nolimits_{i=1}^{N} AP(q_i), \tag{25}$$

where q_i is a query, N is the size of query set, and $AP(q_i)$ is the average precision stated as follows:

$$AP(q)@R = \frac{1}{T}\sum_{i}^{R} \mathrm{Pre}(i)\times\delta(i) \tag{26}$$

where $\mathrm{Pre}(i)$ is the precision of the top i retrieved instances from the top R ranking list. $\delta(i)$ is an indicator function, where $\delta(i) = 1$ denotes that the instance at location i is a relevant sample to query, otherwise $\delta(i) = 0$. T represents the number of relevant samples in the top R ranking result.

In addition, we evaluate the precision and recall through the precision-recall curves, which can reveal the performance of cross-modal retrieval remarkably.

4) IMPLEMENTATION DETAILS

In our experiments, the parameters β, μ, λ and γ of the proposed IISPH are selected from {1e-4, 1e-3, 1e-2, 1e-1}. In addition, the importance of modality is set to $\alpha = 0.5$, and $R = 100$ is set for the computation of mAP. In intra-modality similarity construction, the value of k-nearest neighbors is set to 10. In the distance shrinking phase of each modality, we set $\rho = 0.01$, and ξ is set to the mean value of distance matrix of all instances in each modality. We also investigate the performance with different k, i.e., length of hash codes, which varies in 32, 64, and 128 bits.

B. EXPERIMENTAL RESULTS

1) RESULTS ON WIKI

The mAP scores of different methods on Wiki dataset are reported in Table 2. We can observe that our IISPH outperforms slightly CMFH and SMFH on both Task 1 and Task 2, whereas significantly outperforms CVH and SMH. The main reason is that our method combines the benefits of matrix factorization and supervised label information. Furthermore, with the increasing of hash bits, the performance of our IISPH continuously increases, which can be attributed to its ability to better preserve intra-modality and inter-modality similarities with longer hash bits. The mAP scores of SMFH and CMFH show a similar trend. However, CVH and SMH both degrade slightly. Moreover, the mAP performance of Task 2 is even higher than Task 1. This is because the text representations model the semantic of object better than visual features.

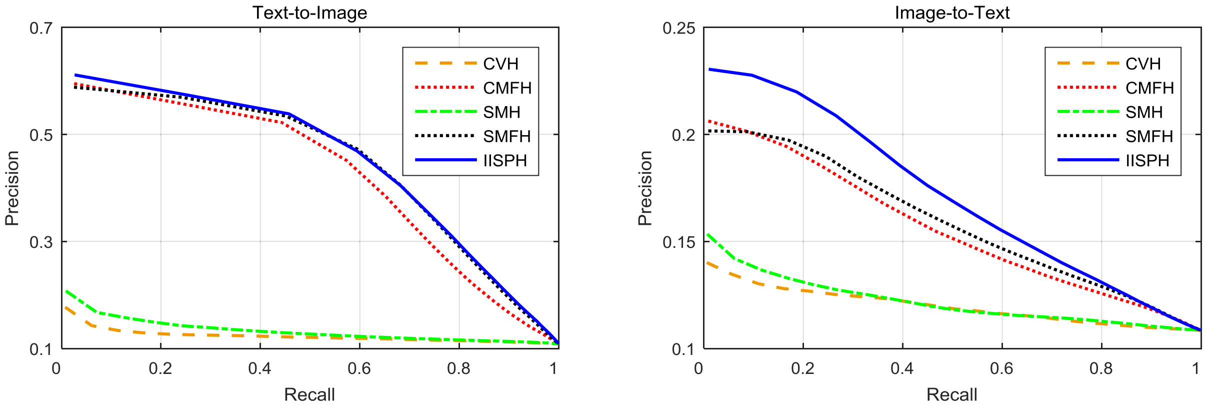
The precision-recall curves of different methods on the Wiki dataset are plotted in Fig. 2. As shown in this figure, the performance keeps consistent with the results of mAP scores, which illustrates that our IISPH consistently outperforms the competitors.

2) RESULTS ON PASCAL

Similar performance gains are observed on the Pascal VOC dataset as shown in Table 2, especially in the task of Text to

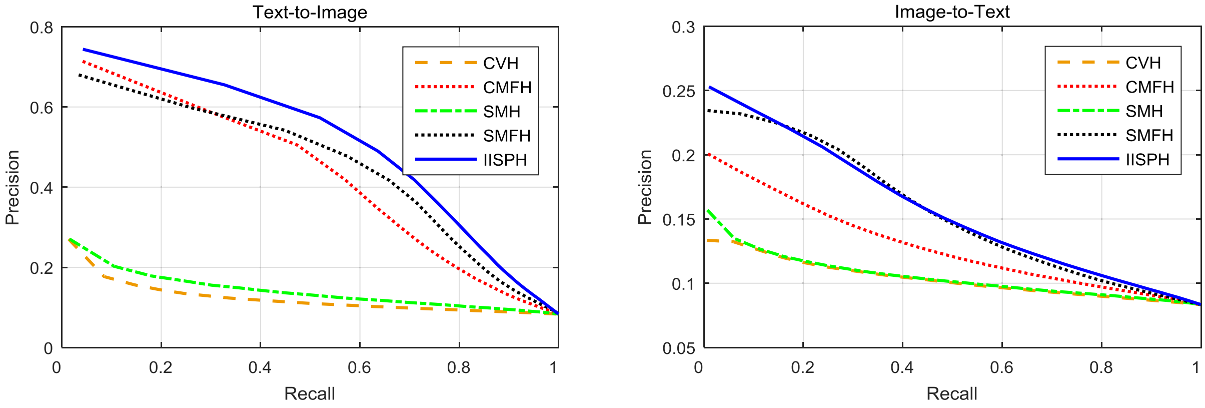
TABLE 2. MAP results @ top 100 on the wiki, pascal voc, and mir flickr dataset with different hash code length.

Task	Methods	WIKI			PASCAL			MIR Flickr		
		32 bits	64 bits	128 bits	32 bits	64 bits	128 bits	32 bits	64 bits	128 bits
Image to Text	CVH	0.1638	0.1531	0.1513	0.1709	0.1695	0.1761	0.6090	0.6066	0.5964
	SMH	0.1975	0.1851	0.1851	0.2158	0.2044	0.1960	0.6230	0.6107	0.6027
	CMFH	0.2320	0.2382	0.2351	0.2282	0.2367	0.2427	0.6216	0.6148	0.6100
	SMFH	0.2290	0.2386	0.2417	0.2510	0.2708	0.2794	0.5253	0.5786	0.6028
	IISPH	**0.2644**	**0.2590**	**0.2665**	**0.2791**	**0.2990**	**0.3043**	**0.6466**	**0.6351**	**0.6258**
Text to Image	CVH	0.2147	0.1829	0.1810	0.2666	0.2994	0.3332	0.6141	0.6103	0.6063
	SMH	0.1945	0.1932	0.1825	0.2779	0.2178	0.1739	0.6304	0.6149	0.6032
	CMFH	0.6027	0.6175	0.6250	0.6912	0.6770	0.6581	0.7018	0.7241	0.7409
	SMFH	0.6057	0.6185	0.6311	0.6695	0.7435	0.7718	0.5288	0.5704	0.5887
	IISPH	**0.6284**	**0.6446**	**0.6446**	**0.7511**	**0.7785**	**0.8233**	**0.7325**	**0.7551**	**0.7773**



(a) (b)

FIGURE 2. Precision-recall curves on Wiki dataset when the number of hash bits is 32. (a) Text to image. (b) Image to text.



(a) (b)

FIGURE 3. Precision-recall curves on Pascal VOC dataset when the number of hash bits is 32. (a) Text to image. (b) Image to text.

Image. Compared to CMFH, our IISPH has achieved 8.6%, 14.9%, and 25% improvement for Task 2 when the number of hash bits are 32, 64, and 128, respectively. It demonstrates that IISPH makes a substantial improvement over CMFH. For Task 2, IISPH has achieved superior performance to SMFH for a performance gain of 12%, 4.7%, and 6.6% when the number of hash bits are 32, 64, and 128 respectively. In terms of Task 1, IISPH slightly outperforms SMFH, whereas significantly outperforms CMFH, CVH, and SMH.

Fig. 3 shows the precision-recall curves of all the counterparts and our IISPH. We can observe that, IISPH obtains superior performance to CMFH consistently, and comparable performance to SMFH. An example of Text to Image of various methods on this dataset is shown in Fig. 5. Given a

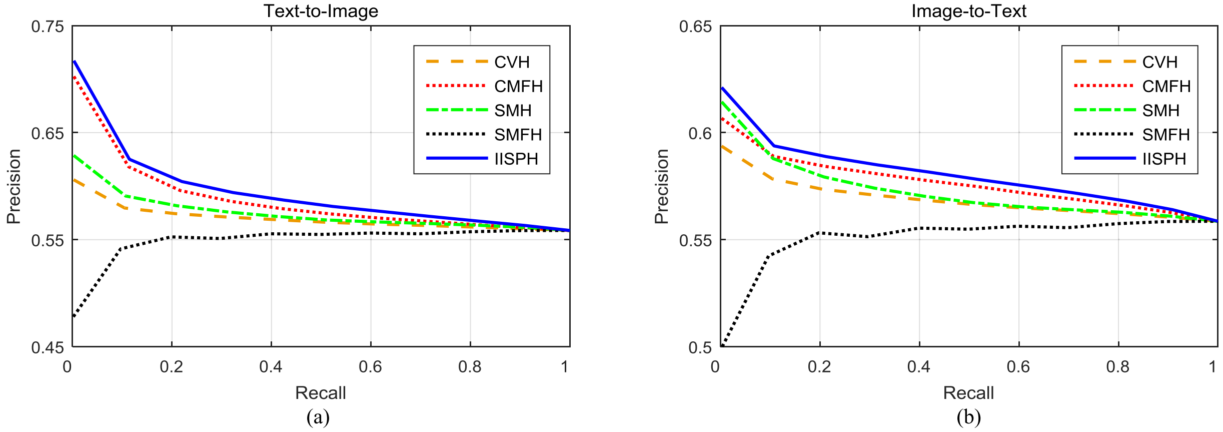


FIGURE 4. Precision-recall curves on MIR Flickr dataset when the number of hash bits is 32. (a) Text to image. (b) Image to text.

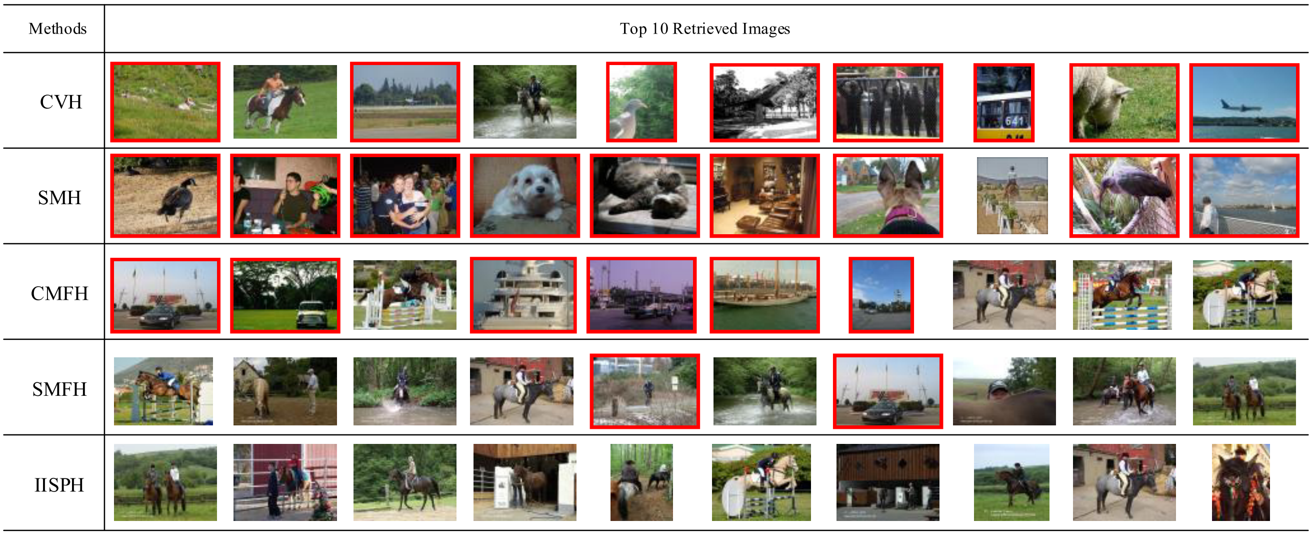


FIGURE 5. An example of cross-modal retrieval task i.e. Text to Image with the query of "horse + person" on the Pascal VOC dataset. Top 10 images retrieved by different methods are presented, where the red border represents an incorrect retrieval result. (Best viewed in color).

text query "horse + person", the top ten images retrieved by CVH, SMH, CMFH, SMFH and IISPH are presented respectively. We can observe that, IISPH returns the perfect results, whereas other methods return with several incorrect images.

3) RESULTS ON MIR FLICKR

Further experiments are conducted on the MIR Flickr dataset. The results of different methods are illustrated in Table 2. In terms of mAP scores, the proposed IISPH outperforms other counterparts, which demonstrates the effectiveness of our method. Specifically, IISPH achieves superior performance to SMFH by about 38%, 32%, and 32% when the hash bits are 32, 64, and 128, respectively for Task 2. The unusual performance of SMFH indicates it is sensitive to input data.

The precision-recall curves of our IISPH and other methods are plotted in Fig. 4. We can observe that, IISPH outperforms other methods in both Task 1 and Task 2 in terms of precision and recall. We further observe that, SMFH achieves the worst performance on this dataset. This phenomenon illustrates that, SMFH is unstable and cannot obtain consistent performance on various datasets. Whereas our IISPH, which encodes the label information into the hashing functions learning procedure, performs consistently well on all different datasets, owing to the learned hashing functions which are more flexible to new coming data.

In summarization, since taking the supervised information into consideration, the performance of our method is superior to that of CMFH. Moreover, from the results of our IISPH and SMFH on the three benchmark datasets, we can observe that, IISPH achieves comparable results to state-of-the-art method SMFH, even better in some cases. In addition, the results of our IISPH are more stable and consistent than that of SMFH. The main reason is that, IISPH embeds the supervised label information into the hashing functions learning, and investigates the local geometric structure in the expected

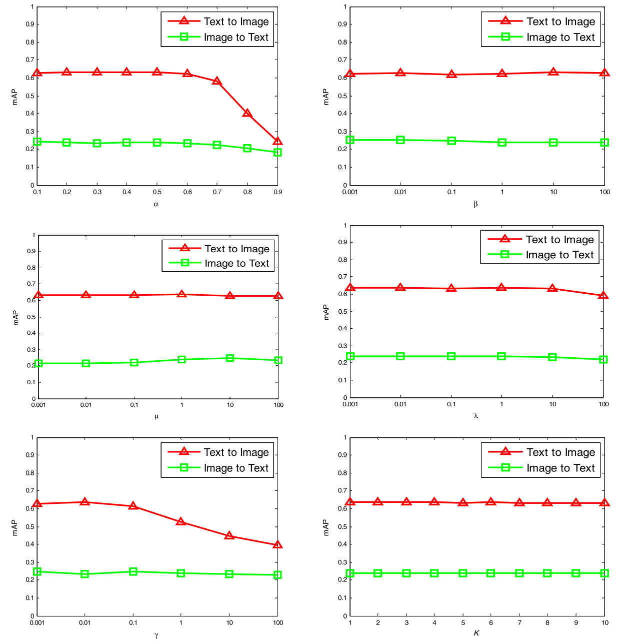
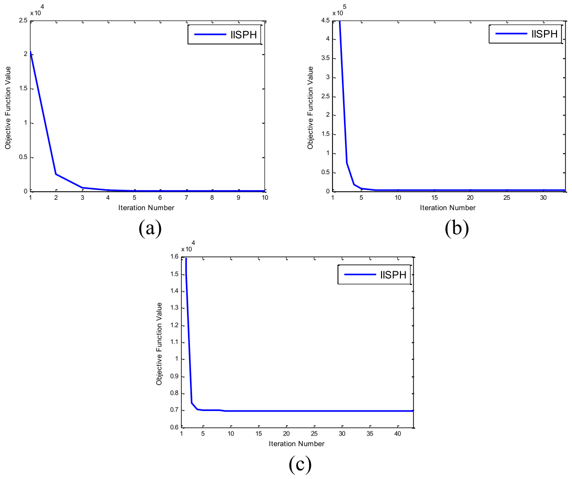


FIGURE 6. Performance variation with respect to parameters α, β, μ, λ, γ, and κ, for both Text to Image and Image to Text tasks on the Wiki dataset when the number of hash bits are 128.



(a) (b)

(c)

FIGURE 7. Convergence curves of the objective function value in (19) on three benchmark datasets when the number of hash bits is 128. (a) Wiki. (b) Pascal VOC. (c) MIR Flickr.

space. All these results demonstrate the effectiveness of the strategies exploited in IISPH.

C. PARAMETER SENSITIVITY ANALYSIS

We also analyze the effects of parameters involved in our IISPH. They are the weight parameter α, trade-off parameters β, μ, λ, γ, and k-nearest neighbors parameter κ. Here, we investigate the performance variation with respect to one parameter in the case of fixing the other parameters. The mAP scores of IISPH with respect to different parameters on the Wiki dataset for both Text to Image and Image to Text tasks are described in Fig. 6, respectively. As shown in Fig. 6, we can see that our IISPH is insensitive to the trade-off parameters β, μ, and λ. We also can conclude that IISPH is not sensitive to the number of k-nearest neighbors κ. As the balance parameter α increases, the weight of text modality 1-α decreases. This leads to the performance of Text to Image task degrades. Since γ controls the regularization term, a too large value will lead to underfitting, which results in the performance degradation of Text to Image.

D. CONVERGENCE

Since the alternative optimization in an iterative manner is used to optimize the objective function, we further investigate the convergence of our algorithm. We conduct experiments on three datasets to recode the objective function value in each iteration, resulting in the convergence curves in Fig. 7. As illustrated in this figure, our algorithm can often converge within 10 iterations, which demonstrates that our IISPH is timing effective for training.

V. CONCLUSION

In this paper, we have proposed an intra-modality and inter-modality similarity preserving hashing for performing cross-modal retrieval. Supervised label information is used to improve the similarity preservation of hash codes. Furthermore, we transformed the view of local consistency preservation from the original space to the expected low-dimensional common space where the local geometric structure is explored with supervised shrinking. In our proposed method, we formulate the label information into the hashing functions learning, which can improve the flexibility of hashing functions for out-of-samples. Thus, our IISPH can achieve superior cross-modal retrieval performance consistently. Experimental results on three benchmark datasets for both cross-modal retrieval tasks have validated the effectiveness of our method, which is superior to state-of-the-art methods. In consideration with the training time, our future work aims at reducing the time complexity of intra-modality similarity.

REFERENCES

[1] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu , "Deep visual-semantic hashing for cross-modal retrieval," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1445–1454.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. 28th IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3156–3164.

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

[4] L. Liu and L. Shao, "Sequential compact code learning for unsupervised image hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2526–2536, Dec. 2016.

[5] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 20th Int. Joint Conf. Artif. Intell.*, Barcelona, Spain, Jul. 2011, pp. 1360–1365.

[6] 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*, Barcelona, Spain, Oct. 2013, pp. 143–152.

[7] 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 Int. Conf. Manage. Data*, New York, NY, USA, Jun. 2013, pp. 785–796.

[8] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th AAAI Conf. Artif. Intell.*, Québec City, QC, Canada, Jul. 2014, pp. 2177–2183.

[9] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2083–2090.

[10] H. Liu, R. Ji, Y. Wu, and G. Hua, "Supervised matrix factorization for cross-modality hashing," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, New York, NY, USA, Jul. 2016, pp. 1767–1773.

[11] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.

[12] D. Rafailidis and F. Crestani, "Cluster-based joint matrix factorization hashing for cross-modal retrieval," in *Proc. 39th Int. ACM Conf. Res. Develop. Inf. Retr.*, Pisa, Italy, Jul. 2016, pp. 781–784.

[13] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Jun. 2016.

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

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

[16] M. Shao and Y. Fu, "Cross-modality feature learning through generic hierarchical hyperlingual-words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 451–463, Feb. 2016.

[17] X. Peng, H. Tang, L. Zhang, Z. Yi, and S. Xiao, "A unified framework for representation-based subspace clustering of out-of-sample and large-scale data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2499–2512, Dec. 2016.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods.," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[19] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

[20] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[21] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin, "Cross-modal subspace learning via pairwise constraints," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5543–5556, Dec. 2015.

[22] K. Wang, W. Wang, R. He, L. Wang, and T. Tan, "Multi-modal subspace learning with joint graph regularization for cross-modal retrieval," in *Proc. 2nd IAPR Asian Conf. Pattern Recognit.*, Naha, Japan, Nov. 2013, pp. 236–240.

[23] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[24] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. 23rd Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2009, pp. 1753–1760.

[25] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *Proc. 26th Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1376–1384.

[26] 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.

[27] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[28] G. Wen and L. Jiang, "Clustering-based locally linear embedding," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Taipei, Taiwan, Oct. 2006, pp. 4192–4196.

[29] D. Xu and S. Yan, "Semi-supervised bilinear subspace learning," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1671–1676, Jul. 2009.

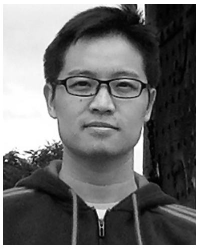
[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[31] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1145–1158, Jun. 2012.

[32] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, Vancouver, BC, Canada, Oct. 2008, pp. 39–43.

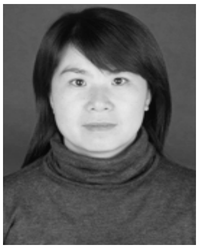
[33] Z. Lin, G. Ding, J. Han, and J. Wang, "Cross-view retrieval via probability-based semantics-preserving hashing," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4342–4355, Dec. 2017.

ZHIKUI CHEN (SM'18) received the B.S. degree from the Department of Mathematics and Computer Science from Chongqing Normal University, China, and the M.S. degree in mechanics and the Ph.D. degree in digital signal processing from Chongqing University, China, in 1993 and 1998, respectively. He is currently a Full Professor with the Dalian University of Technology, China, where he is also leading the Institute of Ubiquitous Network and Computing. His research interests are big data processing, mobile cloud computing, ubiquitous network and its computing.

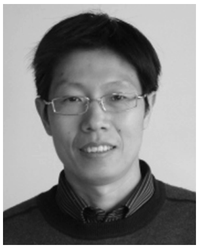


FANGMING ZHONG received the B.S. and M.S degrees in software engineering from the Dalian University of Technology, Dalian, China, in 2012 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Software Technology. His research interests include multimodal learning, cross-modal retrieval, and subspace learning.

GEYONG MIN received the B.Sc. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 1995, and the Ph.D. degree in computing science from the University of Glasgow, Glasgow, U.K., in 2003. He is currently a Professor of high-performance computing and networking with the Department of Mathematics and Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, U.K. His research interests include next-generation Internet, wireless communication, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering.



YONGLIN LENG received the B.S. and M.S. degrees in information science and technology from Bohai University in 2003 and 2009, respectively. She is currently pursuing the Ph.D. degree with the School of Software Technology, Dalian University of Technology, Dalian, China. Her research interests mainly focus on the storage and index of RDF graph data.



YIMING YING received the B.S. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1997 and 2002, respectively. He was a Post-Doctoral Researcher with the City University of Hong Kong, University College London, and the University of Bristol. In 2010, he became a Lecturer (Assistant Professor) in computer science with the University of Exeter. He is currently an Associate Professor with the Department of Mathematics and Statistics, The State University of New York at Albany. His research interests include learning theory, machine learning, and optimization for big data.