Learning Discriminative Binary Codes for Large-scale Cross-modal Retrieval

Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen and Xuelong Li, *Fellow, IEEE*

***Abstract*—Hashing based methods have attracted considerable attention for efficient cross-modal retrieval on large-scale multimedia data. The core problem of cross-modal hashing is how to learn compact binary codes that construct the underlying correlations between heterogeneous features from different modalities. A majority of recent approaches aim at learning hash functions to preserve the pairwise similarities defined by given class labels. However, these methods fail to explicitly explore the discriminative property of class labels during hash function learning. In addition, they usually discard the discrete constraints imposed on the to-be-learned binary codes, and compromise to solve a relaxed problem with quantization to obtain the approximate binary solution. Therefore, the binary codes generated by these methods are suboptimal and less discriminative to different classes. To overcome these drawbacks, we propose a novel cross-modal hashing method, termed Discrete Cross-modal Hashing (DCH), which directly learns discriminative binary codes while retaining the discrete constraints. Specifically, DCH learns modality-specific hash functions for generating unified binary codes, and these binary codes are viewed as representative features for discriminative classification with class labels. An effective discrete optimization algorithm is developed for DCH to jointly learn the modality-specific hash function and the unified binary codes. Extensive experiments on three benchmark datasets highlight the superiority of DCH under various cross-modal scenarios and show its state-of-the-art performance.**

***Index Terms*—Cross-modal retrieval, hashing, discrete optimization, discriminant analysis**

I. Introduction

During the last decade, the tremendous explosion of multimedia data on the Internet has significantly increased the demand for more sophisticated multimedia retrieval technologies. The multimedia data on the Internet usually exist in different media types and come from different data sources, e.g., a blog with texts, images and videos. When searching a topic, it is expected to retrieve a result list simultaneously containing rich information in various media types, which can provide comprehensive description of the topic. However, most existing similarity search methods only apply to a unimodal setting, in which the query item and the retrieved items are in the same modality, e.g., content based image search [1]. These unimodal methods cannot be directly applied to a multi-modal setting because data with different modalities resides in heterogeneous feature spaces.

Actually, *cross-modal retrieval*, which has been widely studied in recent years [2], [3], [4], [5], [6], [7], [8], [9], is becoming more appealing since it supports searching across multi-modal data. Taking multimedia retrieval in image modality and text modality as an example, we can use a query item of one modality (e.g., images) to retrieve relevant items of another modality (e.g., texts). Since heterogeneous data of different modalities resides in different feature spaces, how to model the relationships across these modalities becomes an essential issue that needs to be addressed. In order to eliminate the diversity between different modality features, recent studies are concentrated on latent subspace learning [2], [3], [8], [10], [11] tries to learn a common latent subspace, where the heterogeneity between different modalities is minimized so that the learned features in the subspace can be directly matched. However, one major limitation of these methods is that they are not able to deal with large-scale multi-modal data since the scalability issue is not fully considered.

To tackle the efficiency challenge, several promising approaches [12], [13], [14], [15] introduced unimodal hashing techniques [16], [17], [18], [19], [20], [21], [22] to cross-modal retrieval. To construct the underlying correlations between different modalities, these cross-modal hashing methods find linear projections to embed the heterogeneous data into a common Hamming space, where the multi-modal features are represented by low dimensional binary codes. These methods can handle massive data with both low memory cost and computational cost, because the learned binary codes can be compactly stored and the Hamming distance can be fast calculated with bit-wise XOR operations. In terms of the utilization of label information, the existing cross-modal hashing methods can be grouped into two categories, i.e., unsupervised methods [12], [14], [15], [23], [24], [25] and supervised methods [26], [27], [28], [29], [30], [31]. Specifically, the unsupervised methods generally learn the projections from multi-modal features to binary codes by exploiting the correlations of the training data. On the other hand, the supervised methods extract useful semantic information from the labels of training data, which are further incorporated in the learned binary codes. Generally, label information is helpful for the supervised approaches to achieve better retrieval performance than the unsupervised ones.

Moreover, most supervised cross-modal hashing methods share some common properties. Firstly, they intend to learn binary codes that preserve the inter- and intra-modal similarities, where the pairwise similarities are defined by the given

X. Xu, F. Shen, Y. Yang and H. T. Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (Email: xing.xu@uestc.edu.cn; fumin.shen@gmail.com; dlyyang@gmail.com; shenhengtao@hotmail.com) (Corresponding author: Fumin Shen.)

X. Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China. (Email: xuelong_li@opt.ac.cn)

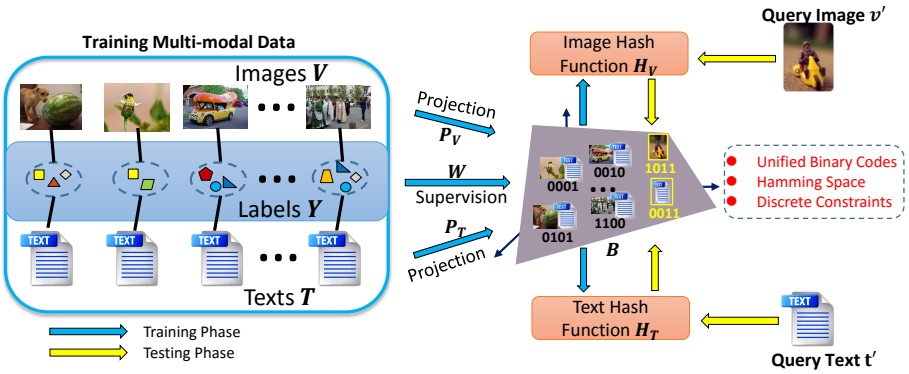


Fig. 1. Framework of the proposed DCH. In the training phase, the multi-modal data of image and text are projected to the common Hamming space, where the unified binary codes of the image-text pairs are further considered as representative features for discriminative classification under the supervision of labels. In the testing phase, given a query sample (image or text), its binary codes are produced via the learned modality-specific hash function.

labels. Secondly, to simplify the discrete optimization problem (which is typically NP-hard) involved in the binary code learning, they choose to relax the original discrete learning problem into a continuous learning problem, and achieve the approximate binary solution via quantization. However, these properties have several intrinsic drawbacks that result in less effective binary codes for these methods. Firstly, preserving the inter- and intra-modal similarities cannot guarantee the learned binary codes to be semantically discriminative. In real scenario, it is natural that multi-modal data with different labels should have discriminative binary codes, while those with the same label have similar binary codes. Secondly, computing the pairwise similarities from given labels inevitably increases the storage and computational cost of these methods. Thirdly, the relaxation scheme may deteriorate the accuracy of the learned binary codes due to the accumulated quantization error.

To address the above challenges, we propose a novel cross-modal hashing method Discrete Cross-modal Hashing (DCH), which learns unified binary codes from multi-modal data more effectively by explicitly capturing the discriminability of labels. Comparing with the existing cross-modal hashing approaches [23], [27] that learn independent binary codes for each modal of one instance, the proposed DCH can efficiently save the storage space of binary codes and reduce the online retrieval time. Fig. 1 depicts the working flow of the proposed DCH. In DCH, the unified binary codes are produced by the modality-specific hash functions that are expected to account the unique structure of individual modality. In particular, inspired by the unimodal hashing methods [19], [32] that use classifiers to learn discriminative binary codes, we consider the to-be-learned unified binary codes as representative features of original multi-modal data for linear classification under the supervision information of labels. Comparing with existing supervised approaches, the main contributions of the proposed DCH can be summarized as follows:

- The proposed DCH jointly learns the unified binary codes and the modality-specific hash functions under the classification framework, hence the learned binary codes are expected to be more discriminative than previous similarity based approaches.
- An efficient discrete optimization algorithm is developed for the proposed DCH to deal with the discrete constraints and to iteratively approximate the optimal binary codes bit by bit in a close form, making the entire learning procedure very efficient to large-scale data.
- The proposed DCH is extensively evaluated on three benchmark datasets with various retrieval scenarios and the results show that it outperforms several state-of-the-art methods.

This paper is an extension and improvement of our previous work presented in [33]. Specifically, 1) we develop an efficient discrete optimization algorithm for DCH in Section III-C, which explicitly takes discrete constraints into the binary code learning process, resulting in codes of higher quality. 2) We discuss the nonlinear case of DCH with kernelized extension in Section III-G, to better capture the underlying nonlinear structure of multi-modal data. 3) We integrate DCH with typical unsupervised hashing methods in a united pipeline, to further bridge the semantic gap between the original representations of multi-modal data. Comprehensive analysis with extensive experiments in Section IV-C demonstrate that the introduced model enhancement schemes above can help to gain performance improvement for the proposed DCH. 4) We design synthetic experiment in Section IV-C to analyze the potential factor that leads DCH to performance degradation in cross-modal retrieval tasks with large code length and provide rational solution that is able to tackle the limitation of DCH in practice.

The remainder of this paper is organized as follows. Section II overviews related work in the field of cross-modal retrieval. Section III elaborates on the proposed DCH, presenting the formula details. Then detailed descriptions of experiments, including experimental settings, results and comprehensive analyses are provided in Section IV. Lastly, the conclusions are made in Section V.

II. Related Work

In this section, we introduce the related work of cross-modal retrieval from two aspects, i.e. latent subspace learning and cross-modal hashing, which are highly related to our work in this paper.

A. Latent Subspace Learning

For the cross-modal retrieval problem, a lot of recent studies are based on latent subspace learning. Canonical Correlation Analysis (CCA) [11] is probably the most popular method that aims to learn a common latent subspace from two modalities where they can be directly matched, while the correlation between them is maximized. CCA has been widely explored and its variants have been proposed for different applications. For examples, Rasiwasia *et al.* [2] applied CCA to learn a maximally correlated subspace for cross-modal retrieval. Other classical works, such as BLM [34] and Generalized Coupled Dictionary Learning (GCDL) [9], are very similar with the CCA since they also learn a common space for cross-modal retrieval. Despite of the above classical methods such as CCA and BLM, other subspace learning approaches leverage the valuable class label information to improve the retrieval performance. Sharma *et al.* [8] proposed a generalized multi-view analysis algorithm to exploit the labels for discriminant latent space learning. Gong *et al.* [35] proposed a multi-view CCA framework for cross-modal retrieval, in which the image and text views are linked under the semantic view of class labels. Kang *et al.* [4] utilized the class labels with a dragging matrix to learn consistent feature representations for two modalities to facilitate the cross-modal matching, and Deng *et al.* [36] proposed a discriminative dictionary learning scheme that augments the correlations among all the modalities by using a common label alignment within the class label space. Moreover, multi-modal deep models such as deep CCA [37], multi-modal auto-encoder [38] and multi-modal restricted Boltzmann machines [39], have also been proposed to address the cross-modal retrieval problem. The target of these methods is to capture the nonlinear transformations from multi-modal data and to construct more powerful subspace in the hidden layers of neural network.

B. Cross-modal Hashing

Recently, an other line of work takes the efficiency issue of cross-modal retrieval into consideration and various hashing approaches have been proposed. According to the discussions in [14], [25], [29], [30], [40], these approaches can be divided into two main categories: unsupervised and supervised hashing methods.

The unsupervised hashing methods generally aims to learn the projections from features to hash codes by exploiting the intra- and inter-view relationship of training data. The motivation in these methods is very similar to the traditional subspace learning methods (e.g., CCA) mentioned above, in that they learn a projected subspace (exactly defined as Hamming space) for cross-modal match by hash codes. For example, inter-media hashing (IMH) was proposed in [12] to learn linear hash functions for mapping features in different views into a common Hamming space by preserving the inter-view and intra-view consistency. Ding *et al.* [15] proposed a method termed CMFH to learn unified hash codes of instances by collective matrix factorization with latent factor model from different modalities. Zhou *et al.* [14] proposed Latent Semantic Sparse Hashing (LSSH) that also generates unified hash codes of instances by first extracting latent semantic features for images and texts respectively with sparse coding and matrix factorization, and then mapping them to a joint abstraction space with quantization. It is worth mention that both CMFH and LSSH utilize matrix factorization technique to mining semantic concepts or latent topics from image or text, which benefits the hash code learning process. Recently, to minimizing the binary quantization errors while preserving data similarities, Irie *et al.* [23] proposed Alternating Co-Quantization (ACQ) that generate binary quantizers for data of one modality with the helpful connections of other modality. ACQ was integrated to CCA and substantially improved its retrieval performance in the Hamming space. Although the results of the above unsupervised methods are promising, hash codes are learned in an unsupervised manner while the available class label information remains unexplored.

On the other hand, the supervised ones further leverage available class label information of training data as supervision to facilitate the hash code learning for performance improvement. The Multi-modal Latent Binary Embedding (MLBE) is proposed in [41] to learn hash functions via a probabilistic model, where hash codes are considered as binary latent factor in a common space, with class labels are utilized to determine both the intra- and inter-modal similarities. Zheng *et al.* [27] proposed co-regularized hashing (CRH) to learn each bit of the hash codes based on a boosted co-regularization framework, and the class labels are used to define the large-margin of each individual modality. Zhang *et al.* [28] proposed semantic correlation maximization (SCM) to seamlessly integrate semantic labels into the hashing learning procedure for large-scale data modeling. Lin *et al.* [29] proposed semantic preserved hashing (SePH) to first generate unified hash codes from the given semantic affinities of training data via minimizing the KL-divergence, and then utilizes kernel logistic regression to learn the non-linear hash functions for each view. Wu *et al.* [42] proposed Sparse multi-modal hashing (SM2H), which utilizes hyper-graph Laplacian sparse coding to learn multi-modal dictionaries while capturing both intra- and inter-modality similarities as a hyper-graph. Inspired the unsupervised CMFH method, Tang *et al.* [30] proposed Supervised Matrix Factorization Hashing (SMFH) that also uses collective matrix factorization to generate unified hash codes to achieve cross-modal search, where the label consistency across different modalities and local geometric consistency in each modality to make the learned hash codes more discriminative. Recently, Jiang *et al.* [31] proposed an end-to-end learning framework with deep neural networks, termed Deep Cross-modal Hashing (DCMH). DCMH integrates feature learning and hash code learning into the same framework, where the to-be-learned hash functions preserve the pairwise similarities defined by class labels.

Generally, a majority of these supervised methods transform the semantic information of given labels into pairwise similarities and then preserving the obtained similarities for the learned hash codes. However, they fail to explicitly capture the category information of given labels, making the learned hash codes less discriminative. Furthermore, these methods typically relax the original discrete learning problem into

a continuous learning problem, which may deteriorate the accuracy of the learned hash codes. In this paper, we proposed a novel DCH method to learn discriminant hash codes more efficiently. Unlike these relaxation-based methods, DCH directly learns the discrete hash codes without relaxation. Given the class label information, DCH casts learning hash codes to a classification framework and jointly learns the modality-specific hash functions and the unified hash codes. The discrete optimization algorithm developed for DCH ensures that each bit is learned successively, thus the final hash codes are more effective to preserve the semantic structure of multi-modal data. Experimental results well demonstrate the advantages of the proposed DCH.

III. Proposed Method

In this section, we present the details of the proposed DCH. To simplify the presentation, we first focus on hash learning for bimodal data (i.e. image and text). Without loss of generality, it can be easily extended to cases with more modalities.

A. Problem Formulation

Suppose that we have a set of n training instances with image-text pairs, denoted by $\mathcal{X} = \{x_i\}_{i=1}^n$. For each instance $x_i = (\mathbf{v}_i, \mathbf{t}_i)$, $\mathbf{v}_i \in \mathbb{R}^f$ is the image feature vector and $\mathbf{t}_i \in \mathbb{R}^d$ is the text feature vector. In addition to feature vectors, class labels $\mathbf{y}_i \in \mathbb{R}^c$ are also available for x_i, where c is the total number of categories, and $y_{ki} = 1$ if x_i belongs to class k and 0 otherwise. Moreover, we respectively denote $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^n \in \mathbb{R}^{f\times n}$, $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^n \in \mathbb{R}^{d\times n}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^{c\times n}$ as the image feature matrix, text feature matrix and class label matrix of all the training instances in $\mathcal{X}$. Without loss of generality, we assume that the feature vectors in $\mathbf{V}$ and $\mathbf{T}$ are both zero-centered, *i.e.* $\sum_{i=1}^n \mathbf{v}_i = \mathbf{0}$ and $\sum_{i=1}^n \mathbf{t}_i = \mathbf{0}$.

Given such data, the goal of DCH is to learn binary codes matrix $\mathbf{B} = \{\mathbf{b}_i\}_{i=1}^n \in \mathbb{R}^{L\times n}$ for training instances in $\mathcal{X}$ and modal-specific hash functions $H_V(\mathbf{v}) = sgn(\mathbf{P}_V^\top \mathbf{v})$ and $H_T(\mathbf{t}) = sgn(\mathbf{P}_T^\top \mathbf{t})$ for image and text modalities, respectively. Here $\mathbf{b}_i \in \{-1, 1\}^L$ is the unified L-bits binary codes vector for instance x_i; $\mathbf{P}_V \in \mathbb{R}^{f\times L}$ and $\mathbf{P}_T \in \mathbb{R}^{d\times L}$ are the linear projection matrices that map the original features of $\mathbf{v}$ and $\mathbf{t}$ to low-dimensional latent spaces, respectively. The sign function $sgn(\cdot)$ outputs +1 for positive numbers and -1 otherwise.

B. Learning Discriminative Cross-modal Binary Codes

In this work, we consider that one instance x_i consisting of features $\mathbf{v}_i$ and $\mathbf{t}_i$ from two modalities has unified binary code $\mathbf{b}_i$, where $\mathbf{b}_i$ is desired to bridge the semantic gap between original features of different modalities. As illustrated in Fig. 1, the original features of image and text modalities are first projected to a common Hamming space. Thus, we assume that the projected features $\mathbf{P}_V^\top \mathbf{v}_i$ and $\mathbf{P}_T^\top \mathbf{t}_i$ of two modalities can produce the same instance $\mathbf{b}_i$ in the common Hamming space, which can be denoted as

$$\mathbf{P}_V^\top \mathbf{v}_i \to \mathbf{b}_i; \qquad \mathbf{P}_T^\top \mathbf{t}_i \to \mathbf{b}_i. \tag{1}$$

Moreover, in order to clearly distinguish the binary codes of different categories, we also expect the binary codes to be representative features of original multi-modal data for discriminative classification. The motivation is that if the binary codes are discriminative enough, they are supposed to be easily classified according to their label information [43]. Therefore, given the class label vector $\mathbf{y}_i$ for x_i , the label vector of the binary feature should be easily predicted by a linear classifier $\mathbf{W} \in \mathbb{R}^{L\times c}$, as

$$\mathbf{y}_i = \mathbf{W}^\top \mathbf{b}_i,\ i = 1, ..., n. \tag{2}$$

Given n training instances, we formulate the binary code learning problem by minimizing the errors of both the projection and classification procedures as follows:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{W},\mathbf{P}_V,\mathbf{P}_T} & \sum_{i=1}^n \|\mathbf{y}_i - \mathbf{W}^\top \mathbf{b}_i\|_2^2 + \mu_V \sum_{i=1}^n \|\mathbf{b}_i - \mathbf{P}_V^\top \mathbf{v}_i\|_2^2 \\ & + \mu_T \sum_{i=1}^n \|\mathbf{b}_i - \mathbf{P}_T^\top \mathbf{t}_i\|_2^2 + \lambda \|\mathbf{W}\|_F^2, \quad s.t. \ \ \mathbf{b}_i \in \{-1, 1\}^L, \end{aligned} \tag{3}$$

where μ_V and μ_T are the penalty parameters, $\|\cdot\|_2^2$ is the L_2 norm for vectors and $\|\cdot\|_F^2$ represents Frobenius norm for matrices, and λ is the regularization parameter. Eq. 3 jointly learns the projection matrices $\mathbf{P}_V$ and $\mathbf{P}_T$, the unified binary codes $\mathbf{B}$ and the linear classifier $\mathbf{W}$. The three terms in Eq. 3 play different roles in the joint learning procedure. Specifically, the first term expects discriminative binary codes for supervised classification, and the second and third terms ensure that the modal-specific hash functions are simultaneously optimized to fit the binary codes with less errors.

C. Optimization Algorithm

Previous works such as [14], [26] simply relax the binary constraints of $\mathbf{b}_i$ to be continuous and cast generating $\mathbf{b}_i$ to directly learning the projection matrices $\mathbf{P}_V$ and $\mathbf{P}_T$ with quantization. However, quantization error may be inevitably accumulated in this approximate solution, resulting in less effective hash functions and binary codes of low quality. To overcome this drawback, in this work we directly learn binary codes with discrete optimization by keeping the binary constraints in the learning procedure.

First, to make the derivation more intuitive, Eq. 3 can be rewritten in matrix form as follows:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{W},\mathbf{P}_V,\mathbf{P}_T} & \|\mathbf{Y} - \mathbf{W}^\top \mathbf{B}\|_F^2 + \mu_V \|\mathbf{B} - \mathbf{P}_V^\top \mathbf{V}\|_F^2 \\ & + \mu_T \|\mathbf{B} - \mathbf{P}_T^\top \mathbf{T}\|_F^2 + \lambda \|\mathbf{W}\|_F^2, \quad s.t. \ \ \mathbf{B} \in \{-1, 1\}^{L\times n}. \end{aligned} \tag{4}$$

The optimization problem in Eq. 4 is non-convex with respect to matrices variables $\mathbf{B}$, $\mathbf{W}$, $\mathbf{P}_V$ and $\mathbf{P}_T$ and difficult to solve. However, it is convex with respect to any one of the four variable while fixing the other three. Therefore, Eq. 4 can be solved by an iterative framework with the following steps until convergence is reached.

Step1: Learn projection matrices $\mathbf{P}_V$ and $\mathbf{P}_T$ by fixing other variables, the problem in Eq. 4 becomes

$$\begin{aligned} & \min_{\mathbf{P}_V} \mu_V \|\mathbf{B} - \mathbf{P}_V^\top \mathbf{V}\|_F^2, \\ & \min_{\mathbf{P}_T} \mu_T \|\mathbf{B} - \mathbf{P}_T^\top \mathbf{T}\|_F^2, \end{aligned} \tag{5}$$

respectively, which can be computed by regression as:

$$\begin{aligned}\mathbf{P}_V &= (\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}\mathbf{B}^\top,\\ \mathbf{P}_T &= (\mathbf{T}\mathbf{T}^\top)^{-1}\mathbf{T}\mathbf{B}^\top.\end{aligned} \tag{6}$$

Step2: Learn classification matrix $\mathbf{W}$ by fixing other variables, the problem in Eq. 4 becomes

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{W}^\top\mathbf{B}\|_F^2 + \lambda\|\mathbf{W}\|_F^2,\ s.t.\ \mathbf{B} \in \{-1,1\}^{L\times n}, \tag{7}$$

which is a simple regularized least squares problem. Thus the close-form solution of $\mathbf{W}$ can be derived as:

$$\mathbf{W} = (\mathbf{B}\mathbf{B}^\top + \lambda\mathbf{I})^{-1}\mathbf{B}\mathbf{Y}^\top. \tag{8}$$

Step3: Learn binary codes matrix $\mathbf{B}$ by fixing other variables, the problem in Eq. 4 becomes:

$$\begin{aligned}&\min_{\mathbf{B}} \|\mathbf{Y} - \mathbf{W}^\top\mathbf{B}\|_F^2 + \mu_V\|\mathbf{B} - \mathbf{P}_V^\top\mathbf{V}\|_F^2 + \mu_T\|\mathbf{B} - \mathbf{P}_T^\top\mathbf{T}\|_F^2,\\ &\quad s.t. \quad \mathbf{B} \in \{-1,1\}^{L\times n}.\end{aligned} \tag{9}$$

The problem in Eq. 9 is NP-hard for directly optimizing the optimal binary codes $\mathbf{B}$. However, similar to [43], $\mathbf{B}$ can be iteratively learned one bit at a time. In other words, the close-form solution of each row of $\mathbf{B}$ can be obtained successively by fixing the other rows. Specifically, the Eq. 9 can be rewritten as follows by expanding each item in Eq. 4:

$$\begin{aligned}&\min_{\mathbf{B}} \|\mathbf{Y}\|_F^2 - 2Tr(\mathbf{Y}^\top\mathbf{W}^\top\mathbf{B}) + \|\mathbf{W}^\top\mathbf{B}\|_F^2\\ &\quad + \mu_V(\|\mathbf{B}\|_F^2 - 2Tr(\mathbf{B}^\top\mathbf{P}_V^\top\mathbf{V}) + \|\mathbf{P}_V^\top\mathbf{V}\|_F^2)\\ &\quad + \mu_T(\|\mathbf{B}\|_F^2 - 2Tr(\mathbf{B}^\top\mathbf{P}_T^\top\mathbf{T}) + \|\mathbf{P}_T^\top\mathbf{T}\|_F^2),\\ &\quad s.t. \quad \mathbf{B} \in \{-1,1\}^{L\times n}.\end{aligned} \tag{10}$$

Note that $Tr(\mathbf{B}^\top\mathbf{B}) = Ln$, then Eq. 10 can be simplified as

$$\min_{\mathbf{B}} \|\mathbf{W}^\top\mathbf{B}\|_F^2 - 2Tr(\mathbf{B}^\top\mathbf{Q}),\ s.t.\ \mathbf{B} \in \{-1,1\}^{L\times n}, \tag{11}$$

where $\mathbf{Q} = \mathbf{W}\mathbf{Y} + \mu_V\mathbf{P}_V\mathbf{V} + \mu_T\mathbf{P}_T\mathbf{T}$ and $Tr(\cdot)$ is the trace norm. It can be seen that the $\mathbf{Q}$ incorporates the information of both image and text features. Fortunately, the form of Eq. 11 is the same as in [43]. Therefore, we can directly leverage the discrete cyclic coordinate descent (DCC) approach proposed in [43] to learn each row (bit) of $\mathbf{B}$ iteratively.

In particular, we denote $\mathbf{z}^\top$ as the l-th row of $\mathbf{B}$ and $\mathbf{B}'$ as the matrix of $\mathbf{B}$ excluding $\mathbf{z}$. Similarly, let $\mathbf{q}^\top$ be the l-th row of $\mathbf{Q}$, $\mathbf{u}^\top$ be the l-th row of $\mathbf{W}$, and $\mathbf{W}'$ be the matrix of $\mathbf{W}$ excluding $\mathbf{u}$. Then the optimal solution of $\mathbf{z}$ can be achieved by the similar derivations in [43], as:

$$\mathbf{z} = sgn(\mathbf{q} - \mathbf{B}'\mathbf{W}'). \tag{12}$$

It is easy to see that in Eq. 12 each bit $\mathbf{z}$ of the training instances is computed depending on the pre-learned $L-1$ bits $\mathbf{B}'$ of the training instances, $\mathbf{q}$ and $\mathbf{W}'$. Thus, it indicates each bit is iteratively updated and the correlations of different modalities can be incorporated in the final codes matrix $\mathbf{B}$. We call the proposed method which discretely learns discriminative binary codes for cross-modal data as *Discrete Cross-modal Hashing* (DCH). In summary, the optimization procedure of DCH is listed in Algorithm 1.

Algorithm 1 Discrete Cross-modal Hashing

Input: Matrices $\mathbf{V}$, $\mathbf{T}$, $\mathbf{Y}$ for image features, text features, class labels respectively; code length L; model parameters λ, μ_V, μ_T;

1: Normalize each column of $\mathbf{V}$ and $\mathbf{T}$ by L_2 norm with zero mean.
2: Initialize $\mathbf{P}_V$, $\mathbf{P}_T$, $\mathbf{W}$ as random matrix respectively, and initialize $\mathbf{B}$ as a $\{-1,1\}^{L\times n}$ matrix randomly.
3: **repeat**
4: Compute $\mathbf{P}_V$, $\mathbf{P}_T$ according to Eq. 6 in Step1.
5: Calculate $\mathbf{W}$ using Eq. 8 in Step2.
6: Learn $\mathbf{B}$ according to the descriptions in Step3.
7: **until** Objective function of Eq. 4 converges.

Output: Binary codes matrix $\mathbf{B}$; modality-specific projection matrices $\mathbf{P}_V$ and $\mathbf{P}_T$.

D. Generating Hash Codes for Queries

Given a new query instance $\mathbf{x}'$, generating its binary codes $\mathbf{b}'$ depends on its components. When $\mathbf{x}'$ contains data of only one modality, it is straightforward to predict its unified binary codes via the modality-specific hash function. When $\mathbf{x}'$ contains data of both two modalities, its unified binary codes are determined by merging the predicted binary codes from different modalities. Thus, the binary codes generation scheme for $\mathbf{x}'$ includes the following two situations:

Only one modality. In this case, $\mathbf{x}' = (\mathbf{v}',)$ or $\mathbf{x}' = (\mathbf{t}',)$. For $\mathbf{x}'$, we directly compute its binary codes $\mathbf{b}'$ as $\mathbf{b}' = sgn(\mathbf{P}_V\mathbf{v}')$ or $\mathbf{b}' = sgn(\mathbf{P}_T\mathbf{t}')$.

Two modalities. In this case, $\mathbf{x}' = (\mathbf{v}',\mathbf{t}')$. For DCH, we add up the results computed by the hash functions of two modalities and generate $\mathbf{b}'$ as $\mathbf{b}' = sgn(\mathbf{P}_V\mathbf{v}' + \mathbf{P}_T\mathbf{t}')$.

E. Complexity Analysis

We discuss the computational complexity of the proposed DCH. In the training phase, the time consuming of each iteration including updating the projection matrices $\mathbf{P}_V$ and $\mathbf{P}_T$, the classifier matrix $\mathbf{W}$ and the unified binary code matrix $\mathbf{B}$. Typically, solving Eq. 6, Eq. 8 and Eq. 12 requires $O(r^2Ln)$, $O(L^2cn)$ and $O(rcLn)$, respectively. Therefore, the time complexity of each iteration is $O((r^2L + L^2c + rcL)nT)$, where $r = \max\{f,d\}$ and T is the number of iterations. It can be observed that the training time is learn to the training set size. Besides, in the experiments part, we will show that DCH usually only needs few iterations (T is very small) to achieve the best modal parameters. Once the training stage is done, the time and space complexities for generating binary codes for a new query are both $O(rL)$ in the query stage, which is extremely efficient. In general, DCH is scalable for large-scale datasets with most existing cross-modal hashing methods and efficient for encoding new query.

F. Extension of Multi-modalities

The proposed DCH can readily be extended to cases of three or more number of modalities. Suppose the training instances $\mathcal{X}$ consists of Z modality data matrices, which are denoted by

X_m, $m = 1,2,...,M$. Then the extension of DCH in Eq. 4 can be formulated as

$$\min_{\mathbf{B},\mathbf{W},\mathbf{P}_m} \|\mathbf{Y} - \mathbf{W}^\top \mathbf{B}\|^2 + \sum_m \mu_m \|\mathbf{B} - \mathbf{P}_m^\top \mathbf{X}_m\|^2 \quad (13)$$
$$+ \lambda \|\mathbf{W}\|^2, \quad s.t. \quad \mathbf{B} \in \{-1,1\}^{L\times n}.$$

where $\{\mathbf{P}_m\}_{m=1}^M$ are the projection matrices for each modality, $\{\mu_m\}_{m=1}^M$ are penalty parameters that balance the contributions of each modality. Derivations of alternating updating rules for $\mathbf{B}$ and $\mathbf{W}$ and each of $\{\mathbf{P}_m\}_{m=1}^M$ are straightforward.

G. Extension of Nonlinear Embedding

The DCH generates linear hash functions as the original features of different modalities are mapped into the common Hamming space via linear projection matrices. To better capture the nonlinear structure underlying the original features, we can adopt non-linear kernel functions to generate kernel features, which essentially represent non-linear projections from original features to binary codes. In our experiments, we choose the typical Gaussian RBF kernel for the nonlinear embedding. Note that, any kernel satisfying the Mercer's condition can be used in our kernelized extension. Specifically, for an instance $x = (\mathbf{v},\mathbf{t})$ with image-text pair, its kernel features $\phi(\mathbf{v})$ and $\phi(\mathbf{t})$ are obtained by the RBF kernel mapping. In particular, $\phi(\mathbf{v}) = [\exp(\|\mathbf{v} - \mathbf{s}_1\|^2/\sigma),...,\exp(\|\mathbf{v} - \mathbf{s}_p\|^2/\sigma)]^\top$, where $\{\mathbf{s}_j\}_{j=1}^p$ are the randomly selected p anchor samples from the training instances $\{\mathbf{v}_i\}_{i=1}^n$ and σ is the kernel width. Similarly, $\phi(\mathbf{t})$ can be obtained as above. Note that the random sampling scheme for computing the kernel features only uses a fraction of the entire training set, making it more efficient and scalable for training. This scheme has been widely adopted in previous studies [19], [19], [25], [29]. Then the non-linear projections of image and text modalities are represented by $\mathbf{P}_V^\top \phi(\mathbf{v})$ and $\mathbf{P}_T^\top \phi(\mathbf{t})$, respectively, which can be fed to DCH to learn hash functions in a kernel space.

IV. Experiment

In this section, we evaluate the retrieval performance and the computational efficiency of the proposed DCH. We first introduce the configuration in our experiments, which includes the datasets, evaluation metrics, parameter settings and comparison methods we used. Then we compare the proposed DCH with several state-of-the-art cross-modal hashing approaches and analyze the results. We further investigate the convergence, computational efficiency and parameter sensitivity of DCH. Finally, we discuss the extension of the proposed DCH, including the nonlinear embedding and the combination with unsupervised cross-modal hashing approaches for performance improvement.

A. Experimental Settings

1) Datasets: To verify the effectiveness of the proposed DCH, we conduct experiments on three benchmark multi-modal datasets, i.e. Wiki [2], MIRFlickr [44] and NUS-WIDE [45], which are widely used in previous studies [14], [15], [23], [25], [29], [30]. All the datasets consist of image and text modalities. Some general statistics of these datasets are given in Table I, and the detailed descriptions for each dataset are as follows:

Wiki consists of 2,866 instances collected from Wikipedia website. It was initially split into a training set of 2,173 instances and a test set of 693 instances. Each instance contains one image and a text document with at least 70 words, and it belongs to one of 10 predefined class labels. For each instance, its image is represented by a 128-dimension SIFT feature vector and its text as a 10-dimension topic vector generated by Latent Dirichlet Allocation (LDA).

MIRFlickr is a real-word dataset originally consisting of 25,000 instances collected from Flickr website, each being an image with its associated textual tags. Each instance is manually annotated with at least one of 24 labels. Similar to the pretreatment in [29], we remove the instances without labels or textual tags appearing less than 20 times. Finally, there are 16,738 instances are left. For each instance, its image is represented by 150D edge histogram, and its text as a 500-dimensional feature vector generated from PCA on its index tagging vector. Here we take 5% of the dataset as the query set and the rest as the training set and retrieval database.

NUS-WIDE is also a real-world image dataset originally containing 269,648 instances. Each instance has an image with its associated textual tags and one of the 81 manually annotated concept labels. Since some of the labels are scarce, following [14], [27], [29], we select the top 10 most frequent labels and thus 186,577 images are left. Furthermore, we select the top 1,000 most frequent tags for these images. Finally, for each instance, its image is represented by a 500D SIFT feature vector and its text as an index vector of selected top 1,000 tags. Here we take 1% of the dataset as the query set and the rest as the training set and retrieval database.

TABLE I
General statistics of three datasets.

	Wiki	MIRFlickr	NUS-WIDE
Training Set Size	2,173	15,902	184,711
Retrieval Set Size	2,173	15,902	184,711
Query Set Size	693	836	1,866
Num. of Labels	10	24	10
Dim. of Image Feature	128	150	500
Dim. of Text Feature	100	500	1,000

2) Baselines and Implementation Details: We compared the proposed DCH with several recently published state-of-the-art cross-modal hashing methods, including unsupervised ones LSSH [14], CMFH [15], CCA-ACQ [23], and supervised ones CRH [27], SCM [28], SEPH [29]. For most baselines we use the code kindly provided by the respective authors, while implementing the CCA-ACQ as its codes are not publicly available. It is notable that the baselines LSSH, CMFH, CRH and SEPH are too computationally intensive, which require much computational time (usually tens of hours) to learn hash functions on the large NUS-WIDE dataset with all training data. To reduce the computational cost of these methods, following the strategy adopted in literatures [14], [15], on NUS-WIDE, we randomly select 10,000 instances from its retrieval set to form the training set for these methods to learn

hash functions and then utilize the learned hash functions to generate binary codes for all instances in the dataset.

The proposed DCH has three model parameters, μ_V and μ_T which balance the discrimination power of image and text modalities, and λ which regularizes the weights of linear classification. In the following section, we provide empirical analysis on parameter sensitivity, which verifies that DCH can achieve stable and superior performance under a wide range of parameter values. When comparing with baseline methods, we empirically set μ_V and μ_T to be 10^{-5}, which shows good performance on three datasets as discussed below. For the baselines, we use the suggested parameters provided by the authors and report the best results of them. For DCH and the baselines with iterative optimization algorithm for parameter learning, we perform 10 runs for them and take the average performance for comparison. All the experiments are performed on Matlab platform installed in a desktop machine which has 8-core 3.4GHz CPUs with 32GB RAM and 64-bit Windows operating system.

3) Evaluation Metrics: In our experiments, we conduct three typical retrieval tasks: 1) *Txt2Img* that uses text as query to search similar images from text database, 2) *Img2Txt* that uses images as query to search similar texts from image database, and 3) *Img2Img* that uses images as query to search similar images from image database. The former two tasks are for cross-modal retrieval, while the latter one is for unimodal retrieval. In testing phase, the Hamming distance is adopted to measure the similarity between the binary codes of query instance and the ones in image/text database. For the proposed DCH and baselines such as LSSH, CMFH and SEPH, which generate unified binary codes for an instance combining both image and text modalities, we consider the image and text databases are equivalent. For the baselines such as CRH and CCA-ACQ which generate different binary codes for each modality of an instance, we separately construct the image and text database based on the their different binary codes.

To evaluate the proposed DCH and baselines, two widely-used performance measures for multi-modal hashing are adopted, namely *mean average precision* (mAP) and topK-precision. To calculate mAP value, we first compute the *average precision* (AP) of each query. Given a query with a group of G retrieved instances, the AP is defined as:

$$AP = \frac{1}{R}\sum_{r=1}^{G} precision(r)\delta(r), \tag{14}$$

where R is the number of ground-truth relevant instances in G, and $\delta(r) = 1$ if the r-th instance is relevant to query and $\delta(r) = 0$ otherwise. For a query, its relevant instances are defined as those share as least one label with it. Then the AP of all queries are averaged to obtain the final mAP. Regarding the metric of topK-precision, it reflects the change of precision with respect to the number of top-ranked K images/texts presented to the users, which is expressive for multimedia retrieval. For both two metrics, larger value indicates better retrieval performance.

B. Overall Comparison with Baselines

1) Results of Img2Txt and Txt2Img tasks: To give an overall evaluation for the proposed DCH and all baseline methods, we first report their cross-modal retrieval performance in terms of mAP on all datasets in Table II, which includes the performance of Img2Txt task and Txt2Img task under typical code length ranging from 16 bits to 128 bits. Among the six baselines in each panel of Table II, the former three are unsupervised methods and the latter three are supervised ones.

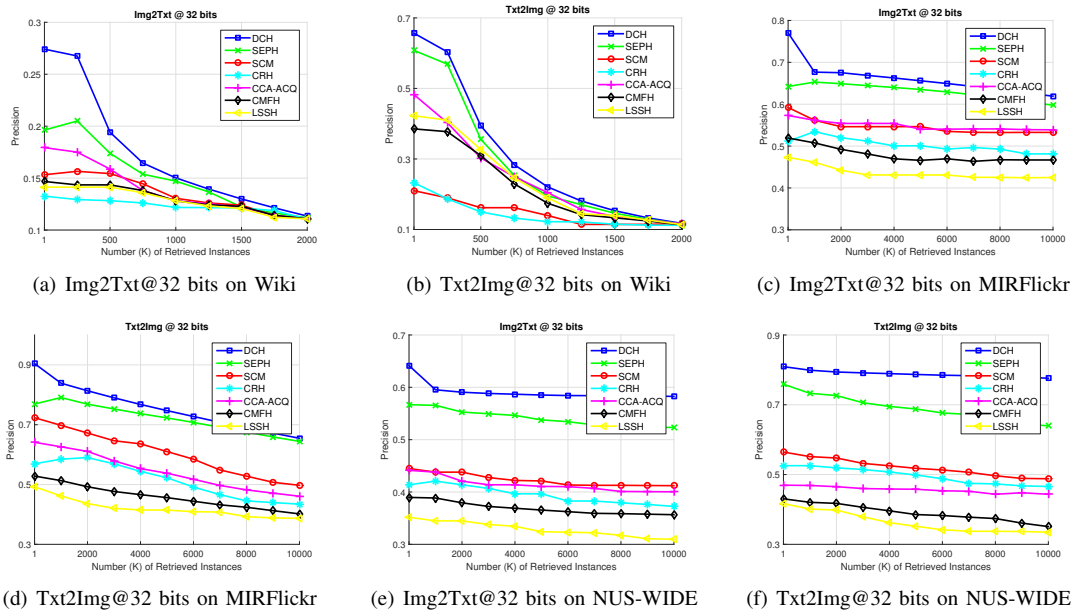
From Table II, we can draw the following observations: 1) The proposed DCH outperforms all the baseline methods with different code lengths on both two retrieval tasks, which well demonstrates the effectiveness of DCH. Specifically, compared with the best baselines, DCH achieves average improvement by 14%, 11%, 12% on Wiki, MIRFlickr and NUS-WIDE respectively. The superiority of DCH can be mainly attributed to its capability to better preserve the discriminability of class labels in the learned unified binary codes, as well as the effectiveness of discrete optimization algorithm developed for DCH that reduces the quantization error in learning hash functions. 2) CCA-ACQ performs much better than the other unsupervised baselines and the supervised baseline CRH, as it is designed to minimize the quantization errors while preserving the data similarities. Therefore, maintenance of quantization quality is crucial to achieve more compact binary codes, which is also fully addressed in the proposed DCH. Moreover, with the helpful supervision of class labels, the proposed DCH remarkably outperforms CCA-ACQ. 3) Comparing with the supervised baselines, the proposed DCH makes significant performance improvement on all datasets. Note that, the second-best supervised approach of SEPH, which captures the pairwise similarities of different modalities, is inferior to learning discriminative binary codes with discrete constraints as the proposed DCH does.

Generally, the mAP scores of DCH consistently increase when the code length becomes longer, since the binary codes are learned bit by bit and the useful information in the pre-learned bits can be incorporated incrementally when updating the current bit. In addition, it also indicates that DCH can encode more discriminative information by using longer binary codes to improve the retrieval performance. Moreover, on all datasets, the DCH and all the baselines usually obtain higher mAP scores on the Txt2Img task than the Img2Txt task, showing that the Img2Txt is more challenging. A potential reason is that the visual feature (e.g., GIST, edge histogram and SIFT) extracted from the image modality might not be informative to describe the high-level semantic. Nevertheless, DCH still outperforms others on Img2Txt task as it generates unified binary codes for an instance with image-text pairs, thus increasing the performance of Img2Txt task by leveraging the information of text modality. It is worth mentioning that the authors in CCA-ACQ has reported that extracting more sophisticated visual features from Convolutional Neural Network (CNN) can significantly improve the performance of Img2Txt task than using the hand-crafted features (e.g., GIST) on Wiki dataset. Later, we have also conducted additional experiment on DCH using the same features as CCA-ACQ does, and have found that DCH still outperforms CCA-ACQ. For example, DCH gains improvement of Img2Txt task by at least 35% on Wiki dataset with various code lengths.

Next, we set the code length to 32 bits and plot the topK-

TABLE II
CROSS-MODAL RETRIEVAL PERFORMANCE OF DCH AND BASELINES ON ALL DATASETS WITH VARIOUS CODE LENGTHS.

Method / Dataset		Wiki				MIRFlickr				NUS-WIDE			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
Img2Txt	LSSH	0.2083	0.2209	0.2214	0.2207	0.5784	0.5804	0.5797	0.5816	0.3955	0.3975	0.4009	0.4018
	CMFH	0.2074	0.2252	0.2357	0.2414	0.5861	0.5835	0.5844	0.5849	0.4327	0.4284	0.4257	0.4236
	CCA-ACQ	0.2672	0.2689	0.2852	0.2679	0.6093	0.6197	0.6002	0.5867	0.4677	0.4554	0.4417	0.4326
	CRH	0.1976	0.1960	0.1978	0.1939	0.5826	0.5745	0.5726	0.5718	0.4536	0.4579	0.4696	0.4713
	SCM	0.2150	0.2330	0.2437	0.2591	0.6069	0.6324	0.6435	0.6476	0.4910	0.5005	0.5006	0.5029
	SEPH	0.2712	0.2947	0.3058	0.3128	0.6541	0.6751	0.6769	0.6803	0.5497	0.5570	0.5603	0.5674
	DCH	**0.3317**	**0.3686**	**0.3762**	**0.3748**	**0.6589**	**0.6801**	**0.6970**	**0.6993**	**0.5789**	**0.5985**	**0.5886**	**0.5775**
Txt2Img	LSSH	0.5021	0.5208	0.5277	0.5325	0.5798	0.5727	0.5732	0.5722	0.4363	0.4299	0.4303	0.4233
	CMFH	0.4874	0.5117	0.5253	0.5354	0.5937	0.5919	0.5931	0.5919	0.4710	0.4611	0.4578	0.4541
	CCA-ACQ	0.5464	0.5612	0.5599	0.5605	0.6145	0.6111	0.6013	0.5865	0.5130	0.4851	0.4776	0.4691
	CRH	0.2629	0.2614	0.2623	0.2615	0.5944	0.5913	0.5838	0.5811	0.4673	0.4714	0.4733	0.4777
	SCM	0.2130	0.2359	0.2472	0.2563	0.6121	0.6190	0.6276	0.6314	0.4618	0.4675	0.4690	0.4709
	SEPH	0.6305	0.6557	0.6626	0.6682	0.7183	0.7249	0.7287	0.7325	0.6415	0.6502	0.6591	0.6672
	DCH	**0.7006**	**0.7087**	**0.7241**	**0.7093**	**0.7381**	**0.7782**	**0.7943**	**0.8127**	**0.7140**	**0.7303**	**0.7162**	**0.6914**



(a) Img2Txt@32 bits on Wiki (b) Txt2Img@32 bits on Wiki (c) Img2Txt@32 bits on MIRFlickr

(d) Txt2Img@32 bits on MIRFlickr (e) Img2Txt@32 bits on NUS-WIDE (f) Txt2Img@32 bits on NUS-WIDE

Fig. 2. The topK-precision curves of DCH and baselines on all datasets.

precision curves of all the methods in Fig. 2 to make a comprehensive contrastive study. From the six sub-figures, it can be seen that DCH always achieves the highest precision than the baselines with the number of retrieved instances (K) changes. This observation is consistent with the overall mAP evaluation in Table II. Especially, when the K is considerably small (e.g., $K \leq 500$ on Wiki and $K \leq 2000$ on MIRFlickr and NUS-WIDE), the precision scores of DCH is much better than the baselines. As in practical retrieval system, users usually pay more attention to the top ranked instances in the retrieved list, therefore, the proposed DCH is superior to the baselines for cross-modal retrieval tasks.

2) Result of Img2Img task: The proposed DCH and baselines can also be applied to unimodal retrieval, as the learned binary codes of them are generated from the common Hamming space of different modalities. In Table III we report their retrieval performance on Img2Img task with different code length ranging from 16 bits to 128 bits on all datasets. Note that as DCH, LSSH and CMFH learn unified binary codes for each training instance, their results on Img2Img task equal to those on Img2Txt task. However, for the other methods that produce different binary codes for different modalities of training instances, their results on Img2Img task vary from those on Img2Txt task. Actually, DCH achieves the best mAP values on Img2Img task compared with the baselines in all cases. Therefore, DCH is effective to naturally leverage the semantic information residing in the text modality data, which improves the unimodal retrieval performance.

C. Comprehensive Analysis on DCH

1) Linear or Nonlinear: In DCH, we learn hash functions from the original multi-modal data of image and text under the linear classification framework. And in Section III-G, we have discussed the nonlinear extension of DCH, which is expected to better capture the underlying nonlinear structure of the multi-modal input data. In this experiment, we extend DCH to the nonlinear case using the RBF kernel mapping for the multi-modal input data, which is denoted as DCH+RBF. Following the suggestions in [19], [29], we empirically take all the training instances as anchors on Wiki, and randomly

TABLE III
UNIMODAL RETRIEVAL (IMG2IMG) PERFORMANCE OF DCH AND BASELINES ON ALL DATASETS WITH VARIOUS CODE LENGTHS.

Method	Wiki				MIRFlickr				NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
LSSH	0.2083	0.2209	0.2214	0.2207	0.5784	0.5804	0.5797	0.5816	0.3955	0.3975	0.4009	0.4018
CMFH	0.2074	0.2252	0.2357	0.2414	0.5861	0.5835	0.5844	0.5849	0.4327	0.4284	0.4257	0.4236
CCA-ACQ	0.1958	0.1920	0.1952	0.1849	0.5563	0.5757	0.5522	0.5477	0.4293	0.4247	0.4184	0.4163
CRH	0.1523	0.1536	0.1558	0.1394	0.5816	0.5615	0.5516	0.5468	0.3901	0.3862	0.3885	0.3924
SCM	0.1657	0.1826	0.1919	0.1862	0.5727	0.5906	0.5876	0.5963	0.4275	0.4288	0.4195	0.4241
SEPH	0.2712	0.2947	0.3058	0.3128	0.6541	0.6751	0.6769	0.6803	0.5497	0.5570	0.5603	0.5674
DCH	**0.3317**	**0.3686**	**0.3762**	**0.3748**	**0.6589**	**0.6801**	**0.6970**	**0.6993**	**0.5789**	**0.5985**	**0.5886**	**0.5775**

TABLE IV
CROSS-MODAL RETRIEVAL PERFORMANCE OF DCH AND DCH+RBF ON ALL DATASETS WITH VARIOUS CODE LENGTHS.

Method / Dataset		Wiki				MIRFlickr				NUS-WIDE			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
Img2Txt	DCH	**0.3317**	**0.3686**	**0.3762**	0.3748	0.6589	0.6801	0.6970	0.6993	**0.5789**	**0.5985**	**0.5886**	**0.5775**
	DCH+RBF	0.3312	0.3643	0.3754	**0.3804**	**0.7041**	**0.7208**	**0.7374**	**0.7655**	0.5159	0.5381	0.4912	0.5024
Txt2Img	DCH	0.7006	0.7087	0.7241	0.7093	**0.7381**	**0.7781**	**0.7943**	**0.8127**	**0.7140**	**0.7303**	**0.7162**	**0.6914**
	DCH+RBF	**0.7245**	**0.7405**	**0.7596**	**0.7457**	0.6855	0.6614	0.6955	0.7323	0.5219	0.5478	0.4972	0.5065

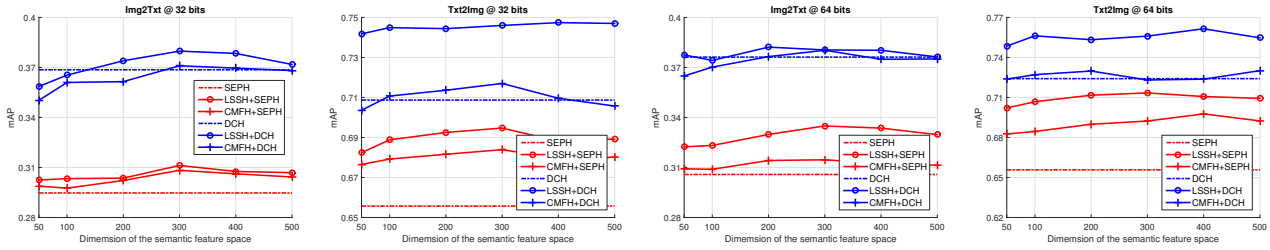
select $p = 500$ on MIRFlickr and NUS-WIDE for RBF kernel mapping, with the parameter σ equaling to 0.01. Then, we analyze the effects of nonlinear embedding by comparing the retrieval performance of DCH+RBF and DCH on all datasets.

Table IV shows their retrieval performance in terms of mAP scores on cross-modal retrieval with various code lengths. Actually, we can observe that using nonlinear embedding may not ensure to achieve better results. Indeed, several factors, such as the extracted features of different modalities, the size of dataset and the specifical retrieval task, may affect the retrieval performance. Firstly, for the Img2Txt task, DCH+RBF outperforms DCH on MIRFlickr dataset, while performing comparably even worse than DCH on Wiki and NUS-WIDE datasets. It indicates that on one hand the nonlinear embedding adopted in DCH+RBF is helpful when the image feature is less representative with low dimension (e.g., 128D edge histogram on MIRFlickr dataset); on the other hand, the nonlinear embedding may harm the capability of the high-dimensional image features such as GIST and SIFT on the other two datasets, which degrades the retrieval performance. Secondly, for the Txt2Img task, DCH+RBF outperforms DCH on Wiki dataset, while being inferior to DCH on MIRFlcikr and NUS-WIDE datasets. Similarly, the nonlinear embedding well explores the underlying nonlinearity of text features with low dimension (e.g., 10D LDA feature on Wikidataset), however, it is not suitable to be applied to the high-dimensional text features such as the frequency histogram used in the other two datasets. Thirdly, on the largest NUS-WIDE dataset, DCH-RBF is seriously inferior to DCH on both two tasks. The reason is that the number of training instances is sufficiently enough for DCH to obtain acceptable performance in the linear case, while using the nonlinear embedding results in overfitting for parameter learning.

2) Integrating DCH with Unsupervised Baselines: In previous experiments, the proposed DCH is applied to the original representations of image and text modalities. However, the semantic gap between the original representations (e.g. visual features and text features) may be quite large, which may lead DCH to fail to capture the common latent information, hence resulting in less effective binary codes. We observe that several unsupervised hashing methods, such as LSSH and CMFH, have shown that explicitly modeling the original representations of both image and text modalities in a joint semantic space before learning hash functions can achieve better results for cross-modal retrieval. Motivated by this observation, we try to further improve the retrieval performance of DCH by combining it with these unsupervised hash methods.

In particular, we first apply LSSH and CMFH on the original representations to extract the semantic features of image and text modalities (in a joint semantic space), then these semantic features are considered as the input data of DCH. These two integration schemes are called LSSH+DCH and CMFH+DCH, respectively. Moreover, we also take the second-best supervised hashing method SEPH as the counterpart, and integrate it with LSSH and CMFH as done for DCH. Similarly, the derived two schemes are respectively denoted as LSSH+SEPH and CMFH+SEPH. We conduct experiment on Wiki dataset, vary the dimension of the joint semantic space extracted by LSSH and CMFH from 50 to 500, and generate binary code length of 32 bits and 64 bits in DCH and SEPH.

Fig. 3 shows the cross-modal retrieval performance of the integration schemes of DCH and SEPH with various dimension of semantic space. From Fig. 3, we can observe that: 1) For SEPH, its integration schemes LSSH+SEPH and CMFH+SEPH consistently achieve higher retrieval performance with different dimensional semantic space on two retrieval tasks; for DCH, its integration schemes (LSSH+DCH and CMFH+DCH) improve DCH when the dimension of the semantic space is considerably higher (e.g. 200D $\sim$ 400D). Therefore, it indicates that using the semantic features extracted from LSSH and CMFH indeed helps to capture the common latent information from the original representations of different modalities, facilitating DCH to learn more effective binary codes for retrieval. 2) Compared with the results on the retrieval task of Img2Txt, the improvements achieved by the two integrations schemes of both DCH and SEPH on Txt2Img are more remarkable. The reason is that the semantic features of the text modalities extracted by LSSH



(a) Img2Txt @ 32 bits (b) Txt2Img @ 32 bits (c) Img2Txt @ 64 bits (d) Txt2Img @ 64 bits

Fig. 3. Cross-modal retrieval performance of DCH, SEPH, and their integration with unsupervised baselines LSSH and CMFH on Wiki dataset. The dimension of the semantic feature space ranges from 50 to 500, and the binary code length are set to be 32 bits and 64 bits.

and CMFH are more informative than those of the image modalities. It again implies that the Img2Txt task is more challenging on Wiki dataset. 3) The integration schemes of DCH consistently performs significantly better than those of SEPH, which showing the superiority of DCH compared with other supervised methods.

3) Effects of Discrete Optimization Algorithm: In this experiment, we investigate the effects of the discrete optimization algorithm adopted in DCH for learning binary codes without relaxation. Specifically, DCH utilizes a discrete cyclic coordinate decent (DCC) approach that has been originally developed in [43] to approximately solve the NP-hard 0-1 quadratic integer programming problem. We compare DCH with our previous work of Discriminant Cross-modal Hashing (DisCH) [33] that has the same form of objective function (Eq. 4) as DCH. However, DisCH differs from DCH that it discards the discrete constraints and solves a relaxed problem by linear programming. We conduct a simulation experiment to solve a simple version of DCH in Eq. 4 with $\lambda = 0.01$ and $\mu_V = \mu_T = 0$. Specifically, we take one training instance from each class on Wiki, resulting in a toy dataset consisting of 10 instances in 10 classes. We then perform DCH and DisCH on the toy dataset to learn 16 bits binary codes for these instances. As the solutions of both DCH and DisCH depend on the initial values of model parameters, we run five trials for both methods with 10 iterations and record their objective values in each trial. Note that, on the toy dataset, the objective value of the theoretically optimum solution can be calculated by exhausting search with 2^{16} binary patterns for each instance.

Fig. 4 shows the changes of the objective values of DCH and DisCH along the iterations for different trials. It can be seen that: 1) both DCH (blue dotted lines) and DisCH (dark dotted lines) find fixed objective values after several iterations (less than 10 iterations), hence they can converge efficiently; 2) DCH generally has lower objective values (around 0.3) comparing to DisCH (around 0.5), which consistently verifies the advantage of the discrete optimization algorithms in DCH on obtaining more effective binary codes; 3) the objective values of DCH (blue dotted lines) are still much larger than the optimal value (red line, around 0.011) and depend on the initialization of model parameters, indicating that the DCC algorithm in DCH is still an approximated solution and may fall into a local minima. Therefore, in real scenario with larger code length, DCH probably learns less effective binary codes compared to those with smaller code length. It is worth mention that the limitation of DCC algorithm has also been discussed in the latest work of [46] for unimodal hashing learning problem, where a scheme that does not require alternating optimization was developed to enhance the robustness of DCC algorithm with larger code length. Intuitively, this scheme can be extended to the cross-modal hashing learning task in the future work.

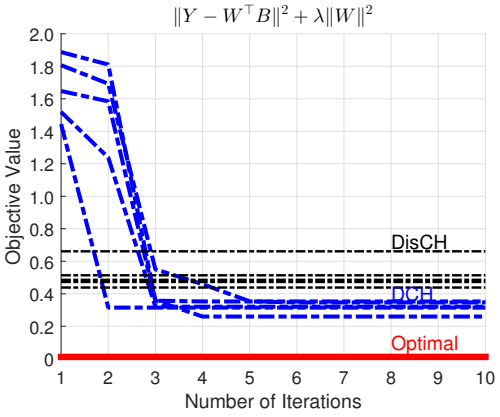
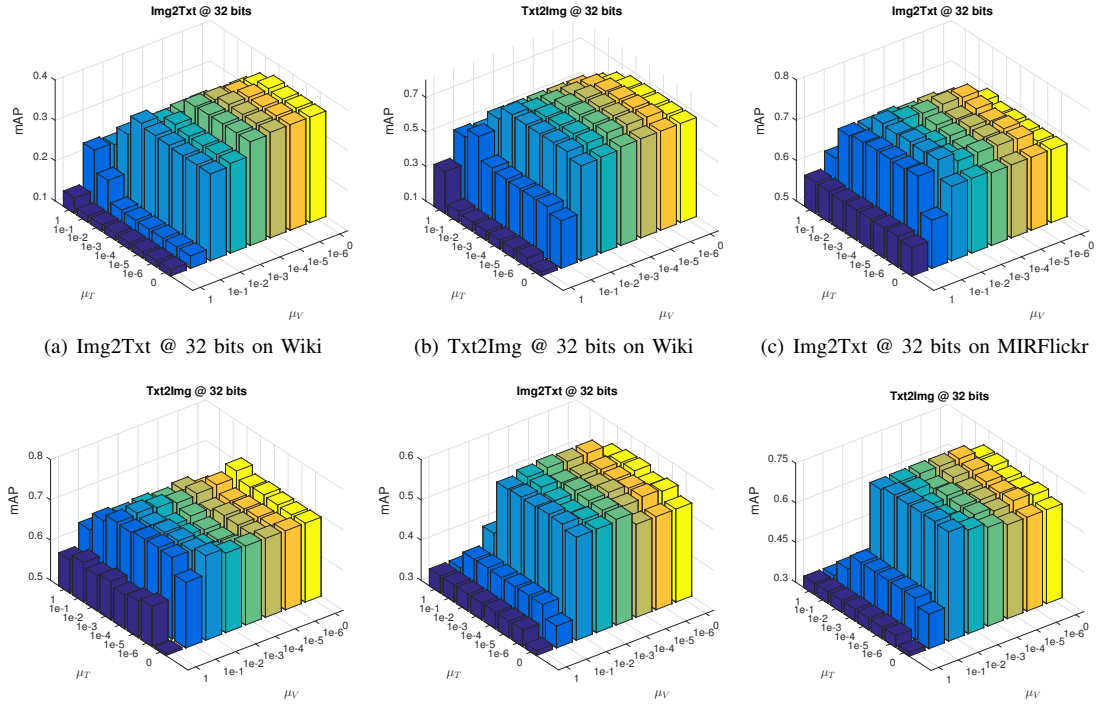


Fig. 4. Changes of objective values obtained by the DCC algorithm adopted in DCH (blue dotted lines) and the relaxation scheme used in DisCH (dark dotted lines), with the theoretical bound as counterpart (red line).

4) Effects of Model Parameters: In previous experiments, we empirically set the penalty parameter μ_V and μ_T in the objective function of DCH (i.e. Eq. 3) as 10^{-5}. As μ_V and μ_T control the error of the learned binary codes and the nonlinear embedding of different modalities, here we analyze their effects on the qualities of the learned binary codes during training. By prefixing the code length as 32 bits, we vary both μ_V and μ_T in $\{0, 10^{-6}, 10^{-5}, ..., 1\}$ and assess their qualities with two retrieval tasks on the training set on all datasets. The evaluation is conducted by changing one parameter (e.g., μ_V) while fixing the other (e.g., μ_T). Note that $\mu_V = 0$ or $\mu_V = 0$ indicates that DCH learns binary codes with the side information of image modality or text modality are neglected, respectively. In addition, when $\mu_V = \mu_T = 0$, DCH learns binary codes only depending on the supervised class label information, where the information of both image and text modalities are ignored.

The six sub-figures in Fig. 5 demonstrate the mAP scores of two retrieval tasks under different settings of μ_V and μ_T



(a) Img2Txt @ 32 bits on Wiki (b) Txt2Img @ 32 bits on Wiki (c) Img2Txt @ 32 bits on MIRFlickr

(d) Txt2Img @ 32 bits on MIRFlickr (e) Img2Txt @ 32 bits on NUS-WIDE (f) Txt2Img @ 32 bits on NUS-WIDE

Fig. 5. Cross-modal retrieval performance of DCH with different values of the model parameters μ_V and μ_T on all datasets.

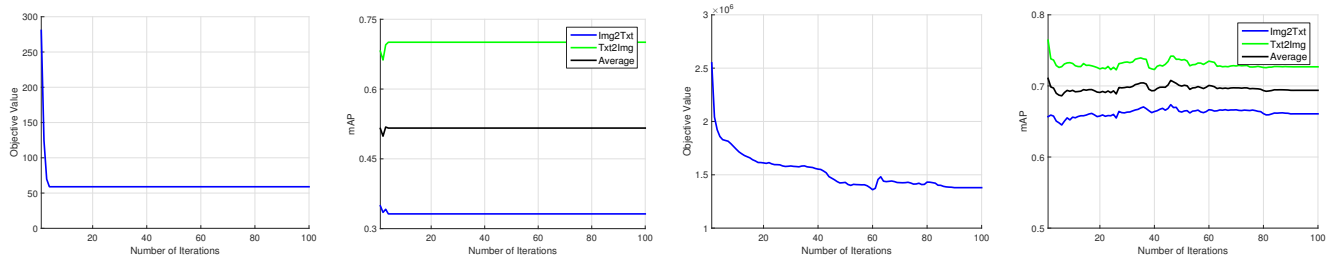
on all datasets. From Fig. 5, we can obtain the following observations. 1) DCH can achieve stable performance under a wide range of values of μ_V and μ_T on all datasets when they are considerably small and comparable, e.g., μ_V and μ_T are from the range of $[10^{-6}, 10^{-2}]$. 2) When μ_V or μ_T is sufficiently large, e.g., μ_V or μ_T equals to 1, the performance deteriorates, as in this case the connection of projection from each modality features to the learned binary codes becomes weak and imprecise. 3) In the case of $\mu_V = 0$ or $\mu_V = 0$, the performance of DCH may be enhanced in one task while deteriorates on the other task. For example, on MIRFlickr in Fig. 5(d), DCH achieves the highest mAP score on Txt2Img task with $\mu_V = 0$ and $\mu_T = 10^{-1}$, while the mAP score on Img2Txt task drops drastically with the same values of μ_V and μ_T. The reason is that when $\mu_V = 0$ the information of the image modality is ignored, then DCH inclines to merely incorporate the side information from the text modality to the learned binary codes. Thus in this case, the learned binary codes of DCH would benefit the Txt2Img task rather than the Img2Txt task. 4) For the special case of $\mu_V = \mu_T = 0$, DCH learns the binary codes with only the supervised class label information, while neglects the information of two modalities. Thus the retrieval performance in this case is inferior to the cases that consider the impact of both two modalities. 5) In practice, to obtain the best retrieval performance on each dataset, we can manually tune the values of μ_V and μ_T in the range of $[10^{-6}, 10^{-2}]$ according to the importance of each modality.

5) Convergence Study: As DCH is solved by an iterative procedure in Algorithm 1, in this part we assess its convergence property by using 32 bits codes on Wiki and MIRFlickr datasets. Indeed, the convergence for the other code lengths are similar as the case of 32 bits codes. Fig. 6(a) and (c) show the convergence curves of the objective value (Eq. 3) with alternating iterations on the two datasets, respectively. And Fig. 6(b) and (d) illustrate the mAP scores of two retrieval tasks and their average with alternating iterations. We can observe that the DCH converges very fast on both datasets. Specifically, for Wiki dataset, it converges within 10 iterations and for MIRFlickr dataset, it converges within 90 iterations. However, in practice, DCH achieves the best mAP score (e.g. "Average" in Fig. 6(b) and (d)) before the objective value converges, i.e. around 5 iterations and 50 iterations for Wiki and MIRFlickr datasets, respectively. Therefore, the overall computational cost of the learning process for DCH is highly efficient due to the possible early-stop before convergence of objective value.

TABLE V
COMPARISON OF TRAINING / TESTING TIME (SECONDS) ON NUS-WIDE DATASET BY VARYING THE SIZE OF TRAINING SET.

Method / Training size	2000	5000	7000	10000	Testing
LSSH	2054	4668	6044	7972	4.366
CMFH	65	146	232	316	0.132
CCA-ACQ	43	92	124	167	0.075
CRH	284	624	866	1234	0.066
SCM	13	15	19	24	0.072
SEPH	1035	2169	3253	4721	0.214
DCH	6	13	19	27	0.032

6) Running Time Comparison: Furthermore, we investigate the running time of the proposed DCH in comparison with the baselines, including the training and testing time of each method. We conduct experiments on NUS-WIDE dataset by



(a) Objective value on Wiki (b) mAP scores on Wiki (c) Objective value on MIRFlickr (d) mAP scores on MIRFlickr

Fig. 6. Convergence analysis. (a) Objective value with alternating iterations on Wiki dataset. (b) mAP scores of two retrieval tasks and their average on Wiki dataset. (c) Objective value with alternating iterations on MIRFlickr dataset. (b) mAP scores of two retrieval tasks and their average on MIRFlickr dataset.

varying the training set size, to analyze the time consumption of model training and testing of each method. Specifically, with the fixed binary code length of 32 bits, we vary the training set size from 2,000 to 10,000 and use the existing query set of 1,866 instances for testing. The record training and testing time costs of each method are shown in Table V. It can be observed that: 1) The training time of all the methods grows linearly with respect to the training set size. 2) LSSH and SEPH cost much more time for training and testing than the other methods. The reason is that before training and testing, LSSH needs to extract sparse representations from different modalities of data and SEPH requires to construct pairwise similarity affinity matrix. These processes increase the computational cost during training. Therefore, they may not be efficient to tackle large-scale data. 3) The proposed DCH performs much more efficiently than the other approaches. It is notable that DCH only takes tens of seconds to training with 10,000 instances, and tens of milliseconds to encode 1,866 query instances, which is extremely fast in practice.

V. Conclusion

In this paper, we proposed a novel supervised hashing method for cross-modal retrieval, named Discrete Cross-modal Hashing (DCH). To achieve discriminative binary codes, DCH was formulated in a linear classification framework, where the class label information was naturally leveraged as supervision during the unified binary code learning process. In this framework, learning the modality-specific hash functions and the unified binary codes was formulated to a joint optimization problem. To tackle the discrete constraints involved in the hash learning problem, we proposed an efficient discrete optimization algorithm, where the optimized binary codes were iteratively obtained bit by bit. We compared DCH with several state-of-the-art baselines on three benchmark datasets, which validated the superiority of DCH on preserving the discrminability in the learned binary codes, and the effectiveness of discrete optimization on reducing the quantization errors. Moreover, we proposed enhancement schemes for DCH to capture the underlying nonlinear structure, and to bridge the semantic gap between the original representations of multi-modal data. These schemes can help DCH to gain further performance improvement in some cases, which are expected to bring several new insights to the community.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Project 61602089, Project 61502081, Project 61572108 Project 61632007 and Project 61472063, the National Thousand-Young-Talents Program of China, and the Fundamental Research Funds for the Central Universities under Project ZYGX2014Z007, Project ZYGX2015J055, Project ZYGX2016KYQD114.

References

[1] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Patt. Reco.*, vol. 40, no. 1, pp. 262–282, 2007.

[2] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. of the 18th ACM Int. Conf. on Multimedia*, 2010, pp. 251–260.

[3] S. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Internation Journal of Computer Vision*, vol. 100, pp. 134–153, 2012.

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

[5] F. Wu, X. Jiang, X. Li, S. Tang, W. Lu, Z. Zhang, and Y. Zhuang, "Cross-modal learning to rank via latent joint representation," *IEEE Trans.s on Image Processing*, vol. 24, no. 5, pp. 1497–1509, 2015.

[6] X. Xu, Y. Yang, A. Shimada, R.-I. Taniguchi, and L. He, "Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts," in *ACM Multimedia*, 2015, pp. 847–850.

[7] Y. Yang, Z. Zha, Y. Gao, X. Zhu, and T. Chua, "Exploiting web images for semantic video indexing via robust sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, 2014.

[8] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 2160–2167.

[9] D. Mandal and S. Biswas, "Generalized coupled dictionary learning approach with applications to cross-modal matching," *IEEE Trans. Image Processing*, vol. 25, no. 8, pp. 3826–3837, 2016.

[10] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *AAAI Conf. on Artificial Intelligence*, 2013.

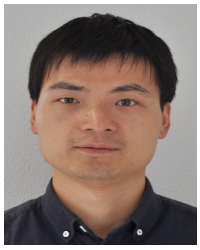
[11] D. Hardoon, S. Szedmak, and J. Shawe-taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computing*, vol. 16, pp. 2639–2664, 2004.

[12] 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. of the 2013 ACM SIGMOD Int. Conf. on Management of Data*, 2013, pp. 785–796.

[13] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. of the Twenty-Second Int. Joint Conf. on Artificial Intelligence*, 2011, pp. 1360–1365.

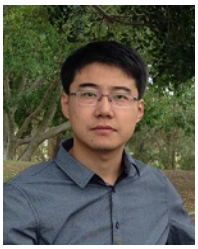
[14] J. Zhou, "Latent Semantic Sparse Hashing for Cross-Modal Similarity Search," *Proc. of the 37th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, pp. 415–424, 2014.

[15] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 2083–2090.
[16] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. of the 20th ACM Symposium on Computational Geometry*, 2004, pp. 253–262.
[17] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in Neural Information Processing Systems, Proc. of the Twenty-Second Annual Conf. on Neural Information Processing Systems*, 2008, pp. 1753–1760.
[18] L. Zhang, Y. Zhang, X. Gu, J. Tang, and Q. Tian, "Scalable similarity search with topology preserving hashing," *IEEE Trans.s on Image Processing*, vol. 23, no. 7, pp. 3025–3039, 2014.
[19] F. Shen, C. Shen, Q. Shi, A. van den Hengel, Z. Tang, and H. T. Shen, "Hashing on nonlinear manifolds," *IEEE Trans.s on Image Processing*, vol. 24, no. 6, pp. 1839–1851, 2015.
[20] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans.s on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2013.
[21] Y. Yang, Y. Luo, W. Chen, F. Shen, J. Shao, and H. T. Shen, "Zero-shot hashing via transferring supervised knowledge," in *Proc. of the ACM Conf. on Multimedia Conf., MM*, 2016, pp. 1286–1295.
[22] D. Zhang, D. Agrawal, G. Chen, and A. K. H. Tung, "Hashfile: An efficient index structure for multimedia data," in *Proc. of the Int. Conf. on Data Engineering, ICDE*, 2011, pp. 1103–1114.
[23] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2015, pp. 1886–1894.
[24] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans.s on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 824–830, 2014.
[25] L. Zhang, Y. Zhang, R. Hong, and Q. Tian, "Full-space local topology extraction for cross-modal retrieval," *IEEE Trans.s on Image Processing*, vol. 24, no. 7, pp. 2212–2224, 2015.
[26] M. Bronstein, A. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 3594–3601.
[27] Y. Zhen and D. Yeung, "Co-regularized hashing for multimodal data," in *Advances in Neural Information Processing Systems*, 2012, pp. 1385–1393.
[28] D. Zhang and W.-J. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *AAAI Conf. on Artificial Intelligence*, 2014.
[29] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.
[30] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans.s on Image Processing*, vol. 25, no. 7, pp. 3157–3166, 2016.
[31] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," *arXiv preprint arXiv:1602.02255*, 2016.
[32] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao, "A fast optimization method for general binary code learning," *IEEE Trans.s on Image Processing*, vol. 25, no. 12, pp. 5610–5621, 2016.
[33] X. Xu, F. Shen, Y. Yang, and H. T. Shen, "Discriminant cross-modal hashing," in *Proc. of the 2016 ACM on Int. Conf. on Multimedia Retrieval*. ACM, 2016, pp. 305–308.
[34] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computing*, vol. 12, no. 6, pp. 1247–1283, 2000.
[35] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. Journal on Computer Vision*, vol. 106, pp. 210–233, 2014.
[36] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Trans.s on Multimedia*, 2015.
[37] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. of the 30th Int. Conf. on Machine Learning*, 2013, pp. 1247–1255.
[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. of the 28th international conference on machine learning*, 2011, pp. 689–696.
[39] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.
[40] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *CoRR*, vol. abs/1606.00185, 2016.
[41] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 940–948.
[42] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, "Sparse multi-modal hashing," *IEEE Trans.s on Multimedia*, vol. 16, no. 2, pp. 427–439, 2014.
[43] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised Discrete Hashing," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 37–45.
[44] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. of the ACM Int. Conf. on Image and Video Retrieval*, 2008, pp. 39–43.
[45] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proc. of the ACM Int. Conf. on Image and Video Retrieval*, 2009, pp. 48:1–48:9.
[46] G. Koutaki, K. Shirai, and M. Ambai, "Fast supervised discrete hashing and its analysis," *CoRR*, vol. abs/1611.10017, 2016.



Xing Xu received the B.E. and M.E. degrees from Huazhong University of Science and Technology, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Japan, in 2015. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, China. His research interests include multimedia information retrieval and pattern recognition.

Fumin Shen received the B.S. from Shandong University in 2007 and the Ph.D. degree from the Nanjing University of Science and Technology, China, in 2014. He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China. His major research interests include computer vision and machine learning, including face recognition, image analysis, hashing methods, and robust statistics with its applications in computer vision.



Yang Yang received the bachelors degree from Jilin University in 2006, the masters degree from Peking University in 2009, and the Ph.D. degree from The University of Queensland, Australia, in 2012, under the supervision of Prof. H. T. Shen and Prof. X. Zhou. He was a Research Fellow with the National University of Singapore from 2012 to 2014. He is currently with the University of Electronic Science and Technology of China.

Heng Tao Shen is a Professor in University of Electronic Science and Technology of China. He obtained his BSc with 1st class Honours and PhD from Department of Computer Science, National University of Singapore in 2000 and 2004 respectively. He then joined the University of Queensland as a Lecturer, Senior Lecturer, Reader, and became a Professor in late 2011. His research interests mainly include Multimedia/Mobile/Web Search, and Big Data Management on spatial, temporal, multimedia and social media databases. Heng Tao has extensively published and served on program committees in most prestigious international publication venues of interests. He received the Chris Wallace Award for outstanding Research Contribution in 2010 conferred by Computing Research and Education Association, Australasia. He is currently an Associate Editor of IEEE Transactions on Knowledge and Data Engineering (TKDE).

Xuelong Li (M'02-SM'07-F'12) is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.