J. Vis. Commun. Image R. 79 (2021) 103256

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Robust multimodal discrete hashing for cross-modal similarity search ☆

Yuzhi Fang *

School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China
College of Information Engineering, Shandong Management University, Jinan 250357, Shandong Province, China

ARTICLE INFO

Keywords:
Hashing
Robust
Cross-modal retrieval
Unsupervised learning

ABSTRACT

Hashing technology improves the search efficiency and reduces the storage space of data. However, building an effective modal with unsupervised cross modal retrieval and generating efficient binary code is still a challenging task, considering of some issues needed to be further discussed and researched for unsupervised multimodal hashing. Most of the existing methods ignore the discrete restriction, and manually or experientially determine the weights of each modality. These limitations may significantly reduce the retrieval accuracy of unsupervised cross-modal hashing methods. To solve these problems, we propose a robust hash modal that can efficiently learn binary code by employing a flexible and noise-resistant $\ell_{2,1}$-loss with nonlinear kernel embedding. In addition, we introduce an intermediate state mapping that facilitate later modal optimization to measure the loss between the hash codes and the intermediate states. Experiments on several public multimedia retrieval datasets validate the superiority of the proposed method from various aspects.

1. Introduction

Hash technology [1] has shown its superiority in multimedia [2–4], computer vision [5–8], machine learning [9–12], information retrieval [13,14] and other fields. It has advantages of low storage and fast retrieval, and makes efficient and fast high-dimensional data retrieval become a research hotspot. How can the hash code effectively preserve the inter-modal and intra-modal similarities of heterogeneous data in original feature space is the key problem of multimodal hashing. In the past decades, many researchers have proposed various hash methods to generate effective hash codes.

Considering the necessity of preserving the intra-modal and inter-modal similarities at the same time, Kumar et al. proposes cross-view hash (CVH) [15], which is used to learn hash function for each view given the set of multi-view training data objects. However, as the number of bits increases, the retrieval performance decreases, because most of the variance in the matrix decomposition process is contained in the first few eigenvectors. Semantics-Preserving Hashing method (SEPH) [16] achieves the conversion between the semantic affinity and probability distribution of the training data, and approximate it by minimizing Kullback Leibler divergence. For any invisible instance, a new probability method is used to determine the unified hash code by using the predicted hash code and its corresponding output probability in the observation view. When semantic tags are used to the replace pairwise similarity, the retrieval performance improves, but the training time complexity increases.

Although gratifying achievements have been made, there still exist some problems and unsupervised multi-modal hashing methods based on matrix decomposition need further consideration. Firstly, it is usually NP-hard problem that the integer restriction of binary coding in hash function learning leads to discrete optimization problems. To avoid the problems of discrete optimization in binary code learning, most existing methods firstly solve the relaxation problem by directly discarding discrete constraints, and then symbolizing the continuous output into binary. The importance of discrete optimization in existing hashing methods is seldom considered. Objectively speaking, discarding the discrete constraints in the objective function may leads to a large amount of error accumulation in the process of representation learning and symbolization of real values. In fact, ignoring the discrete constraints of binary code is one of the main reasons for the decrease of search performance [17]. Secondly, the quality of the approximate solution is usually very low, and the resolution of the hash function generated by iteration is usually degraded due to the accumulated quantization errors, especially when learning long codes. The accumulated ℓ_2-norm errors in matrix factorization are obvious because they are sensitive to the output and their robustness need to be improved. Thirdly, different dimensions of multimodality have different modal weights. However, the weights are often assigned manually. How to effectively obtain the appropriate weight for each modal remains a problem.

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Correspondence to: School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China.
E-mail address: abcdfangyuzhi@163.com.

https://doi.org/10.1016/j.jvcir.2021.103256
Received 29 July 2020; Received in revised form 9 June 2021; Accepted 31 July 2021
Available online 15 August 2021

To address these problems, a novel unsupervised multi-modal hashing method called Robust Multimodal Discrete Hashing for Cross-modal Similarity Search (RMDH) is proposed to learn jointly discrete binary codes as well as robust hash functions. Specifically, in RMDH, we avoid the discrete optimization problem instead of optimizing the binary code directly. At the same time, we notice that the generated binary codes inevitably contain some noises that interfere with the subsequent semantic indexing tasks, and these noises produce errors accumulation. Therefore, in order to suppress the influence of perturbed binary codes and the accumulation of errors, for robust hash function learning, we also integrate the anti-jamming $\ell_{2,1}$ form loss with the embedding of non-linear kernels [18]. In addition, during parameter optimization, we learn the weights for each modal from data distribution, rather than being artificially assigned. The main contributions of RMDH are summarized as follows:

- A robust matrix factorization method is proposed for directly learning the consistent binary code of heterogeneous data, which avoids the quantization errors caused by relaxation during the optimization process and significantly improves the retrieval performance.
- In order to increase the robustness of the model, the $\ell_{2,1}$ loss function is introduced to effectively alleviate the adverse effects of noisy data.
- In order to control the sampling rate between different modalities, in the hash code learning process, the data of each modal will get a reasonable weight, which can better balance the proportion of each modal and facilitate subsequent learning work.

The rest of this paper is organized as follows. The related literature are briefly reviewed in Section 2. Section 3 presents the proposed robust multimodal discrete hashing for cross-modal similarity search, and give the theoretical analysis. Section 4 report the experimental results on three datasets and give comparative analysis of the methods, lastly the conclusions are made in Section 5.

2. Related work

Available supervisory information, such as paired constraints or labels are applied to supervise multi-modal hashing for preserving semantic similarities. Recently, more and more attention have been paid to supervised multi-modal hash methods. These methods try to obtain semantic relevance in the semantic tags of training points, and their effects are higher than unsupervised multi-modal hash methods. Cross-modality similarity-sensitive (CMSSH) hashing proposed by Bronstein et al. [19] attempts to study supervised multimodal hashing. Zhen et al. proposes a multimodal latent binary embedding (MLBE) approach [20], which is a probabilistic model by considering the binary latent factors as hash codes in a common Hamming space. Semantic correlation maximization (SCM) [21] makes full use of semantic label information and uses it for large-scale data retrieval modeling of hash code learning, and greatly reduces the time cost. Multi-modal discriminative binary embedding (MDBE) [22] uses label information to discover the shared structure within heterogeneous data, and keeps the original data structure to learn the differentiability and similarity of hash codes.

Recently, with the improvement of data computing capability, deep learning method has attracted more and more attention because of its significant advantages in studying the internal structure of multi-dimensional data, such as image classification [23–26], speech recognition and synthesis [27,28], object detection [29–31]. With reference to the types of unimodal hashing, general multimodal hashing methods can be divided into two categories: supervised and unsupervised [32]. Supervised learning discrete hash (SLDH) [33], utilizes the deep learning framework to learn the compact binary code for image retrieval. The original features are transformed into binary code by multi-layer network, and the semantic relevance of artificial tags are used to keep the semantic similarity. With the continuous development of deep learning technology, researchers began to use deep neural networks to obtain multimodal hashing. Multimodal hashing methods based on deep learning usually convert heterogeneous features of different modals into isomorphic features of the same modal to generate hash codes for each modal data, and capture the inherent cross-modal correspondence between heterogeneous data by using supervisory information [34–36].

Cross-view hashing (CVH) [15] is a pioneer in the research and supervision of multi-modal hash. In this paper, the author studies the modeling of multi-view training data objects, and maps similar objects in different views to similar codes to achieve cross-view similarity search. In order to solve the needs of large-scale data retrieval, the inter-media hashing (IMH) [37] uses different data types as the medium to study the correlation between multimedia and creatively solve the scalability problem. Considering the problem of time complexity, LCMH [38] is proposed to determine the size of training data in order to achieve scalable index of multimodal search. LCMFH [39] transforms heterogeneous data into potential semantic space, and keeps the generated hash code consistent with the semantic label of the original data, and directly uses the semantic label to guide the hash learning process.

The above-mentioned multimodal retrieval method needs to construct a similarity graph combined with eigenvalue decomposition to learn a hash function. However, the similarity map and eigenvalue decomposition require a lot of time and cost, and with the increase of sample data and the length of the hash code, the mapping effect fluctuates greatly. Pedronette et al. [40] proposes a method to improve the efficiency of image retrieval tasks by using the inherent geometry of the dataset and the data manifold. STMH [41] learns the relationship between different topics and different data sources, determines the relationship between different modalities in the learning semantic space, and directly generates a unified hash code. STMH is much suitable for hash learning, and can get good retrieval performance. UCAL [42] maximizes the correlation between models and introduces a modal classifier to predict the modality of the transformed features. It can be regarded as the statistical regularization of feature transformation to ensure that the transformed features cannot be distinguished statistically. Wang et al. [43] proposes an online cross-modal retrieval method, which effectively alleviates the problem that the batch-based hash learning model can only learn for fixed data and it has poor scalability. In order to learn binary codes with better discrimination and flexibility. Lu et al. [44] proposes to automatically learn the weights of each modal and learn the hash function according to the new data stream, avoiding repeated learning.

3. Robust multimodal discrete hashing

In order to facilitate the following expressions, we reduce the modal types to two categories, i.e. image and text hash code learning. Obviously, it can be easily extended to the case of multimodal types.

3.1. Notations and problem formulation

Suppose that n objects with image–text pairs, can be represented by $O = \{o_i\}_{i=1}^{n}$, $o_i = \left(\mathbf{x}_1^i, \mathbf{x}_2^i\right)$, where $\mathbf{x}_1 \in \mathbb{R}^{d_1}$, $\mathbf{x}_2 \in \mathbb{R}^{d_2}$ are two feature vectors from image and text. d_1 and d_2 are the dimensionalities of the feature vector for each modality, respectively, and usually $d_1 \neq d_2$. The above vectors are expressed in matrix form $\mathbf{X}_1 = [\mathbf{x}_1^1, \mathbf{x}_1^2, \ldots, \mathbf{x}_1^n] \in \mathbb{R}^{d_1 \times n}$, $\mathbf{X}_2 = [\mathbf{x}_2^1, \mathbf{x}_2^2, \ldots, \mathbf{x}_2^n] \in \mathbb{R}^{d_2 \times n}$, respectively. $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n] \in \mathbb{R}^{k \times n}$ is the hash codes that we are trying to learn, where $b_j \in \{0,1\}^{k \times 1}$, k is the length of the hash code. The purpose of RMDH is to learn unified binary codes $\mathbf{B} \in \mathbb{R}^{k \times n}$, for data entities O, in this way, b_i and b_j retain the semantic similarity between o_i and o_j as much as possible.

3.2. Collective factorization for cross-modality hashing

Non-negative matrix factorization (NMF) is a matrix factorization algorithm, which can reduce dimension by learning equivalent low rank vectors. Collective Matrix factorization (CMF) [45] is a method to learn the relationship among multimodal data by matrix factorization, which predicts relationships between entity datasets. Many multimodal hash retrieval methods have used this method successfully [46–48]. We give non-negative data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, the projection vectors from $\mathbf{V}$ to them are denoted by $\mathbf{U}_1$, $\mathbf{U}_2$, whose product could approximate $\mathbf{X}_i(i = 1, 2)$, i.e.,

$$\mathbf{X}_i \approx \mathbf{U}_i\mathbf{V} \tag{1}$$

where $\mathbf{X}_i \in \mathbb{R}^{d_i\times n}$, $\mathbf{U}_i \in \mathbb{R}^{d_i\times k}$, $\mathbf{V} \in \mathbb{R}^{k\times n}$. Most cross-modal hashing problems use the squared Frobenius norm to represent the errors in the above approximation process,

$$\min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{V}} \lambda\|\mathbf{X}_1 - \mathbf{U}_1\mathbf{V}\|_F^2 + (1-\lambda)\|\mathbf{X}_2 - \mathbf{U}_2\mathbf{V}\|_F^2, \tag{2}$$

where $\lambda \in (0, 1)$ is a balance parameter. After the unified representation $\mathbf{V}$ is obtained, the hash code of the image–text pair can be expressed by $sgn(\mathbf{V} - \tilde{\mathbf{V}})$, where $\tilde{\mathbf{V}}$ is the average of the vectors.

We replace $\mathbf{V}$ by $\mathbf{B}$, and have:

$$\min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{B}} \alpha_1^\gamma\|\mathbf{X}_1 - \mathbf{U}_1\mathbf{B}\|_{2,1} + \alpha_2^\gamma\|\mathbf{X}_2 - \mathbf{U}_2\mathbf{B}\|_{2,1}, \tag{3}$$

where α_1^γ, α_2^γ $(\gamma > 1)$ are used to control the weight distribution of the image and text.

Eq. (3) has the following advantages over Eq. (2). Firstly, we introduce $\ell_{2,1}$−norm loss function to enhance the robustness. The $\ell_{2,1}$−norm of a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m] \in \mathbb{R}^{n\times m}$ is defined as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n}\|\mathbf{a}_i\|_2. \tag{4}$$

Secondly, the weight parameters can be learned. For different modality data, we learn different weights so that important modality is given large weight.

3.3. Graph regularization

In this section, we briefly review the details of spectral hashing and introduce the formulation of using local geometric structure. Suppose that we have n data points $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$. The main idea of spectral hashing is to effectively retrieve similar samples with as few bits as possible, and similar samples should also have similar binary codes.

Firstly, we define $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_n]^T \in \{0, 1\}^{k\times n}$, where $\mathbf{b}_i \in \{0, 1\}^{k\times 1}$ is $x_i's$ binary code vector. For similarity measures, we generally agree with the following rules, assuming that two samples are similar under the same similarity measure, they should also have similar binary codes in Hamming space. In order to achieve this goal, we set up the following model:

$$\begin{aligned} &\min_{\mathbf{b}_i} \sum_{i,j} \mathbf{s}_{ij}\|\mathbf{b}_i - \mathbf{b}_j\|^2 \\ &s.t. \quad \mathbf{b}_i \in \{0, 1\}^{k\times 1}, i = 1, 2, \dots, n, \end{aligned} \tag{5}$$

where $\|\mathbf{b}_i - \mathbf{b}_j\|$ computes the Hamming distance between $\mathbf{b}_i$ and $\mathbf{b}_j$. It can be found that the more similar $\mathbf{b}_i$ and $\mathbf{b}_j$ are, the smaller Hamming distance$\|\mathbf{b}_i - \mathbf{b}_j\|$ is. The intra-modal similarity $\mathbf{s}_{ij}$ of two data $\mathbf{x}_i$ and $\mathbf{x}_j$ from the same modality is defined as follows:

$$\mathbf{s}_{ij} = \begin{cases} exp(-\dfrac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}), if\ \mathbf{x}_i \in \mathbf{N}_l(\mathbf{x}_j)\ or\ \mathbf{x}_j \in \mathbf{N}_l(\mathbf{x}_i) \\ 0, \qquad\qquad otherwise, \end{cases} \tag{6}$$

where $\mathbf{N}_l(\cdot)$ is the set of the l nearest neighbors and σ is the bandwidth parameter. Given the dataset $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, the weighted adjacency matrix $\mathbf{S} = (s_{ij})_{i,j=1,2,\dots,n}$ construct the undirected graph. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected, then $s_{ij} > 0$, otherwise, $s_{ij} = 0$. Then, we represent $\mathbf{Z}$ as a diagonal matrix, where the main diagonal element is $z_{ij} = \sum_j s_{ij}$, so we set the Laplacian of the graph as: $\mathbf{L} = \mathbf{Z} - \mathbf{S}$ which represents the graph Laplacian matrix calculated using different strategies based on the internal geometry of the data source. By simplifying the calculation, we get the regularization terms of mixed graphs, and Eq. (5) can be redefined as follows

$$\min_{\mathbf{B}} Tr(\mathbf{B}^T\mathbf{L}\mathbf{B}) \tag{7}$$

3.4. Learning robust discrete spectral hashing

In this section, we introduce the learning of discrete binary code functions for constructing image semantic indexing tasks. As previously assumed, suppose that we have data $\mathbf{X} \in \mathbb{R}^{d\times n}$, d and n represent the feature dimension and the number of samples, respectively. Our goal is to preserve the semantic similarity among data points in the Hamming space and learn effective and efficient binary codes. Then, we denote $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_n] \in \{0, 1\}^{k\times n}$, where the ith column vector $\mathbf{b}_i$ is the k-bits hash codes for $\mathbf{x}_i$. In this work, we choose spectral hashing to represent learning model, together with the Iterative Quantization [6], and obtain the following,

$$\begin{aligned} &\min_{\mathbf{B},\mathbf{F},\mathbf{Q}} \beta\|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2 + \eta Tr(\mathbf{F}\mathbf{L}\mathbf{F}^T), \\ &s.t.\ \mathbf{B} \in \{0, 1\}^{k\times n}, \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_k, \end{aligned} \tag{8}$$

where $\mathbf{Q} \in \mathbb{R}^{k\times k}$ is an orthogonal matrix. By rotating the intermediate representation $\mathbf{F}$, $\mathbf{F}\mathbf{Q}$ can approximate the binary code $\mathbf{B}$ at the right angle. It is worth mentioning that the rotation matrix introduced to prevent the generation of trivial solution is orthogonal.

Now, we summarize our overall function of RMDH as below:

$$\begin{aligned} \min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{B},\mathbf{F},\mathbf{Q}} \mathbf{O} = &\sum_{v=1}^{2}(\alpha_v)^\gamma\|\mathbf{X}_v - \mathbf{U}_v\mathbf{B}\|_{2,1} + \beta\|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2 \\ &+ \eta Tr(\mathbf{F}\mathbf{L}\mathbf{F}^T) \end{aligned} \tag{9}$$

$$s.t.\ \mathbf{B} \in \{-1, 1\}^{k\times n}, \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_k, \alpha_1 + \alpha_2 = 1, \alpha_1 > 0, \alpha_2 > 0.$$

where α_i is the weight coefficient of each modality satisfying $\alpha_1+\alpha_2 = 1$, β and η are the tradeoff parameters of the corresponding terms. $\mathbf{Q}\mathbf{Q}^T$ reduces the redundancy among hash bits.

In the next section, we focus on solving the optimization problems of the proposed model.

3.5. Optimization algorithm

Generally speaking, due to the special structure of $\ell_{2,1}$-norm in Eq. (9), it is more difficult to solve $\ell_{2,1}$-norm than to solve ℓ_2-norm, in order to facilitate later calculations, we write Eq. (9) as follows:

$$\begin{aligned} \mathbf{O} = &\sum_{v=1}^{2}\left\{(\alpha_v)^\gamma Tr\big((\mathbf{X}_v - \mathbf{U}_v\mathbf{B})\mathbf{D}_v(\mathbf{X}_v - \mathbf{U}_v\mathbf{B})^T\big)\right\} \\ &+ \beta\|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2 + \eta Tr(\mathbf{F}\mathbf{L}\mathbf{F}^T) \end{aligned} \tag{10}$$

$$s.t \quad \mathbf{B} \in \{-1, 1\}^{k\times n}, \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_k, \alpha_1 + \alpha_2 = 1, \alpha_1 > 0, \alpha_2 > 0.$$

where $\mathbf{D}_v$ is a diagonal matrix with its ith diagonal element defined as

$$(\mathbf{D}_v)_{ii} = 1/\|\mathbf{x}_v^i - \mathbf{U}_v\mathbf{b}_i\|_2, \tag{11}$$

Next, we focus on solving Eq. (10), by alternately updating $\mathbf{U}_v$, $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{F}$, and $\mathbf{D}_v$.

(1) Fixing $\mathbf{B}$, $\mathbf{F}$, $\mathbf{D}_v$, $\mathbf{Q}$ and α_v, and updating latent factors matrix $\mathbf{U}_v$ for each modality.

We derive the partial derivatives with respect to $\mathbf{U}_v$ as:

$$\frac{\partial\mathbf{O}}{\partial\mathbf{U}_v} = 2(\alpha_v)^\gamma\mathbf{U}_v\mathbf{B}\mathbf{D}_v\mathbf{B}^T - 2(\alpha_v)^\gamma\mathbf{X}_v\mathbf{D}_v\mathbf{B}^T + 2\theta\mathbf{U}_v, \tag{12}$$

Let $\frac{\partial \mathbf{O}}{\partial \mathbf{U}_v} = 0$, we can get closed-form solution

$$\mathbf{U}_v = \mathbf{X}_v\mathbf{D}_v\mathbf{B}^T(\mathbf{B}\mathbf{D}_v\mathbf{B}^T + \theta\mathbf{I}/(\alpha_v)^\gamma)^{-1}. \tag{13}$$

where θ is a parameter introduced to prevent overfitting, and $\mathbf{I}$ is an identity matrix.

(2) Fixing $\mathbf{U}_v$, $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{D}_v$ and α_v, updating $\mathbf{B}$.

In order to update hash code $\mathbf{B}$ conveniently, we transform Eq. (10) into the following form

$$\begin{aligned}\min_B \mathbf{O} = &\sum_{v=1}^{2}\left\{(\alpha_v)^\gamma Tr\big((\mathbf{X}_v - \mathbf{U}_v\mathbf{B})\mathbf{D}_v(\mathbf{X}_v - \mathbf{U}_v\mathbf{B})^T\big)\right\}\\ &+ \beta\|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2\\ &s.t. \quad \mathbf{B} \in \{-1, 1\}^{k\times n}.\end{aligned} \tag{14}$$

Constrained by discretization, it is difficult to directly solving $\mathbf{B}$ in Eq. (14). Inspired by [49], we use the discrete cyclic coordinate descent (DCC) method. The main operation is to fix all other rows, learn one bit each time, so that each row of $\mathbf{B}$ has a display solution. Then, we rewrite Eq. (14) as following:

$$\begin{aligned}\min_{\mathbf{B}} \mathbf{O} = &\sum_{v=1}^{2}\left\{(\alpha_v)^\gamma Tr\big(\mathbf{U}_v\mathbf{B}\mathbf{D}_v\mathbf{B}^T\mathbf{U}_v^T - 2\mathbf{D}_v\mathbf{X}_v^T\mathbf{U}_v\mathbf{B}\big)\right\}\\ &- 2\beta Tr(\mathbf{F}^T\mathbf{Q}^T\mathbf{B})\\ &s.t \quad \mathbf{B} \in \{-1, 1\}^{k\times n}.\end{aligned} \tag{15}$$

The above formula is equivalent to

$$\begin{aligned}\min_{\mathbf{B}} \mathbf{O} = &\sum_{v=1}^{2}\left\{(\alpha_v)^\gamma Tr\big(\mathbf{U}_v\mathbf{B}\mathbf{D}_v\mathbf{B}^T\mathbf{U}_v^T\big\} - Tr(\mathbf{G}_v\mathbf{B})\right)\\ &s.t \quad \mathbf{B} \in \{-1, 1\}^{k\times n}.\end{aligned} \tag{16}$$

where $\mathbf{G}_v = 2\sum_{v=1}^{2}(\alpha_v)^\gamma\mathbf{D}_v\mathbf{X}_v^T\mathbf{U}_v + 2\beta\mathbf{F}^T\mathbf{Q}^T$.

Let $\mathbf{b}$ is $\mathbf{B}'$ jth row, and $\mathbf{B}'$ be the matrix of $\mathbf{B}$ excluding $\mathbf{b}$. Similarly, let $\mathbf{g}_v$ is $\mathbf{G}_v'$ jth column, $\mathbf{G}_v$ be the matrix of $\mathbf{G}_v$ excluding $\mathbf{g}_v$, $\mathbf{u}_v$ is $\mathbf{U}_v$ jth column, and $\mathbf{U}_v'$ be the matrix of $\mathbf{U}_v$ excluding $\mathbf{u}_v$, then we have

$$\begin{aligned}&\sum_{v=1}^{2}\left\{(\alpha_v)^\gamma Tr\big(\mathbf{U}_v\mathbf{B}\mathbf{D}_v\mathbf{B}^T\mathbf{U}_v^T\big)\right\}\\ &= \sum_{v=1}^{2}(\alpha_v)^\gamma\left(2\mathbf{u}_v^T\mathbf{U}_v'\mathbf{B}'\mathbf{D}_v b^T + \mathbf{u}_v^T\mathbf{u}_v\mathbf{b}\mathbf{D}_v\mathbf{b}^T\right) + const.\end{aligned} \tag{17}$$

Similarly, we have

$$\begin{aligned}Tr(\mathbf{G}_v\mathbf{B}) &= \{Tr(\mathbf{G}_v'\mathbf{B}') + \mathbf{g}_v^T\mathbf{b}^T\}\\ &= \mathbf{g}_v^T\mathbf{b}^T + const,\end{aligned} \tag{18}$$

Combining (13)–(15), we obtain an objective function that follows $\mathbf{b}$

$$\begin{aligned}\min_{\mathbf{B}} \mathbf{O} = &\sum_{v=1}^{2}(\alpha_v)^\gamma\left(2\mathbf{u}_v^T\mathbf{U}_v'\mathbf{B}'\mathbf{D}_v b^T + \mathbf{u}_v^T\mathbf{u}_v\mathbf{b}\mathbf{D}_v\mathbf{b}^T - \mathbf{g}_v^T\right)\mathbf{b}^T\\ &s.t \quad \mathbf{B} \in \{-1, 1\}^{k\times n}.\end{aligned} \tag{19}$$

The above formula has the following optimized solution

$$\mathbf{b} = sgn\Big(\mathbf{g}_v^T - \sum_{v=1}^{2}(\alpha_v)^\gamma\left(2\mathbf{u}_v^T\mathbf{U}_v'\mathbf{B}'\mathbf{D}_v\mathbf{b}^T - \mathbf{u}_v^T\mathbf{u}_v\mathbf{b}\mathbf{D}_v\mathbf{b}^T\right)\Big). \tag{20}$$

(3) Fix $\mathbf{U}_v$, $\mathbf{B}$, $\mathbf{Q}$, $\mathbf{D}_v$ and α_v, updating $\mathbf{F}$.

Then the problem can be transformed to

$$\begin{aligned}&\min_{\mathbf{B}\in\{-1,1\}^{k\times n}} \beta\|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2 + \eta Tr(\mathbf{F}\mathbf{L}\mathbf{F}^T)\\ &= \min_{\mathbf{B}\in\{-1,1\}^{k\times n}} \beta Tr\big((\mathbf{B} - \mathbf{Q}\mathbf{F})(\mathbf{B} - \mathbf{Q}\mathbf{F})^T\big) + \eta Tr(\mathbf{F}\mathbf{L}\mathbf{F}^T),\end{aligned} \tag{21}$$

differentiate $\mathbf{F}$ by the above formula and simplify it to get

$$\beta\mathbf{Q}^T\mathbf{Q}\mathbf{F} + \eta\mathbf{F}\mathbf{L} - \beta\mathbf{Q}^T\mathbf{B} = 0. \tag{22}$$

Obviously, Eq. (22) is a Sylvester equation [50], which can be solved by using the lyap function of Matlab.

(4) Fixing $\mathbf{U}_v$, $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{D}_v$, updating $\mathbf{Q}$.

Then obtain

$$\begin{aligned}&\min_{\mathbf{Q}} \|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2\\ &s.t. \quad \mathbf{Q}\mathbf{Q}^T = \mathbf{I}_k.\end{aligned} \tag{23}$$

Theorem 1. *Suppose the singular value decomposition of $\mathbf{B}\mathbf{F}^T$ is $\mathbf{U}\Sigma\mathbf{V}^T$, then the closed form solution of $\mathbf{Q}$ is $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$.*

Proof. We construct the Lagrange function as

$$\mathbf{L}(\mathbf{Q}, \mathbf{A}) = \|\mathbf{B} - \mathbf{Q}\mathbf{F}\|_F^2 + Tr\big(\mathbf{A}(\mathbf{Q}\mathbf{Q}^T - \mathbf{I})\big), \tag{24}$$

which equivalent to

$$\begin{aligned}L(\mathbf{Q}, \mathbf{A}) &= Tr\big((\mathbf{B} - \mathbf{Q}\mathbf{F})(\mathbf{B}^T - \mathbf{F}^T\mathbf{Q}^T)\big)\\ &\quad + Tr\big(\mathbf{A}(\mathbf{Q}\mathbf{Q}^T - \mathbf{I})\big)\\ &= Tr(\mathbf{F}\mathbf{F}^T\mathbf{Q}^T\mathbf{Q}) - 2Tr(\mathbf{Q}\mathbf{F}\mathbf{B}^T)\\ &\quad + Tr(\mathbf{Q}^T\mathbf{A}\mathbf{Q}) + const,\end{aligned} \tag{25}$$

where $\mathbf{A} \in \mathbb{R}^{k\times k}$ is the Lagrange multiplier and is a symmetric matrix. Since $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, Eq. (25) is equivalent to

$$L(\mathbf{Q}, \mathbf{A}) = -2Tr(\mathbf{Q}\mathbf{F}\mathbf{B}^T) + Tr(\mathbf{Q}^T\mathbf{A}\mathbf{Q}) + const. \tag{26}$$

We set the derivative of $L(\mathbf{Q}, \mathbf{A})$ with respect to $\mathbf{Q}$ to zero, and obtain

$$\frac{\partial L(\mathbf{Q}, \mathbf{A})}{\partial \mathbf{Q}} = -\mathbf{B}\mathbf{F}^T + \mathbf{A}\mathbf{Q} = 0. \tag{27}$$

Then, we have

$$\mathbf{Q} = \mathbf{A}^{-1}\mathbf{B}\mathbf{F}^T. \tag{28}$$

equivalent to $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}$, we get

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{F}^T\mathbf{F}\mathbf{B}^T\mathbf{A}^{-1} = \mathbf{I}. \tag{29}$$

Thus, we have

$$\mathbf{A} = \left(\mathbf{B}\mathbf{F}^T(\mathbf{B}\mathbf{F}^T)^T\right)^{1/2}. \tag{30}$$

By substituting (30) in (28), we have

$$\begin{aligned}\mathbf{Q} &= \left(\mathbf{B}\mathbf{F}^T(\mathbf{B}\mathbf{F}^T)^T\right)^{-1/2}\mathbf{B}\mathbf{F}^T\\ &= \left(\mathbf{U}\Sigma\mathbf{V}^T\mathbf{V}\Sigma\mathbf{U}^T\right)^{-1/2}\mathbf{U}\Sigma\mathbf{V}^T\\ &= \mathbf{U}\mathbf{V}^T.\end{aligned} \tag{31}$$

End of the proof.

(5) Fix $\mathbf{U}_v$; $\mathbf{B}$; $\mathbf{Q}$; $\mathbf{F}$ and $\mathbf{D}_v$, updating α_v.

Obviously, we rewrite O as

$$\begin{aligned}&\min_{\alpha_v} O = \sum_{v=1}^{2}(\alpha_v)^\gamma\mathbf{H}_v\\ &s.t. \sum_{v=1}^{2}\alpha_v = 1, \alpha_v > 0,\end{aligned} \tag{32}$$

where $\mathbf{H}_v = Tr\big((\mathbf{X}_v - \mathbf{U}_v\mathbf{B})\mathbf{D}_v(\mathbf{X}_v - \mathbf{U}_v\mathbf{B})^T\big)$.

Similarly, we solve the above optimization problem by constructing the Lagrange function

$$\min_{\alpha_v} O = \sum_{v=1}^{2}(\alpha_v)^\gamma\mathbf{H}_v - \delta\Big(\sum_{v=1}^{2}\alpha_v - 1\Big). \tag{33}$$

We take the derivative of α_v in Eq. (33) and set its derivative to 0, combining it with the constrain $\Sigma_{v=1}^{2}\alpha_v = 1$, we obtain

$$\alpha_v = \frac{(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}}{\sum_{v=1}^{2}(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}}. \tag{34}$$

Algorithm 1 Robust Multimodal Discrete Hashing (RMDH)

Input: feature matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, parameters β, η.
Output: hash codes $\mathbf{B}$.
Initialize:
1.Initializing $\mathbf{U}_1, \mathbf{U}_2, \mathbf{B}, \mathbf{F}, \mathbf{Q}$ randomly respectively.
2.Initializing $\mathbf{D}_v = \mathbf{I}_n$ for each modality.
Repeat:
1.Fixing **B**, **Q**, **F**, $\mathbf{D}_v$ and α_v , updating $\mathbf{U}_1, \mathbf{U}_2$ by Eq.(13);
2.Fixing $\mathbf{U}_v$, **Q**, **F**, $\mathbf{D}_v$ and α_v, updating **B** by Eq.(20);
3.Fixing $\mathbf{U}_v$, **B**, **Q**, $\mathbf{D}_v$ and α_v, updating **F** by Eq.(22);
4.Fixing $\mathbf{U}_v$, **B**, **F**, $\mathbf{D}_v$ and α_v, updating **Q** by Eq.(31);
5.Fixing $\mathbf{U}_v$, **B**, **Q**, **F** and $\mathbf{D}_v$, updating α_v by Eq.(34);
6.Fixing $\mathbf{U}_v$, **B**, **Q**, **F** and α_v, updating $\mathbf{D}_v$ by Eq.(11);
until:convergence

(6) Fixing $\mathbf{U}_v$, **B**, **Q**, **F** and α_v, updating $\mathbf{D}_v$ for each modality by Eq. (11).

Through the six steps above, we update $\mathbf{U}_v$, **B**, **F**, **Q**, α_v, $\mathbf{D}_v$ one by one. The above iterative process is repeated until the objective function converged. The RMDH approach is summarized in Algorithm 1.

3.6. Times complexity

The main time consumption of the proposed RMDH is the matrix factorization, The time complexity of solving Eqs. (13), (20), (22), (31), (34) and (11) is $\mathcal{O}(k^2 d_v + k^2 n + k^3), \mathcal{O}(d_v k n), \mathcal{O}(k^2 n), \mathcal{O}(k^2 n), \mathcal{O}(d_v^2 n + d_v k n)$ and $\mathcal{O}(d_v k n)$, respectively, where d_v is the dimension of the modality, k is the length of hash code, n is the number of training samples. Normally, k and d_v is much smaller than n, so we obtain the time complexity of RMDH is linear to the size of the data, which is scalable.

3.7. Out-of-sample extension

After learning the hash code of the training set, it is very important to solve the problem of out-of-sample expansion of RMDH. However, in order to obtain the hash code of the unseen sample to retrain the RMDH is not worth the gain. Therefore, we designed a model for online learning of new sample hash codes.

Given a query sample $\mathbf{X}_v$, we focus on solving the following model optimization problems:

$$\min_{\mathbf{B}} \mathbf{O} = \sum_{v=1}^{n_t} (\alpha_v)^\gamma \|\mathbf{X}_v - \mathbf{U}_v \mathbf{B}\|_{2,1} \tag{35}$$

$$s.t. \mathbf{B} \in \{-1, 1\}^{k \times n}, \sum_{v=1}^{n_t} \alpha_v = 1, \alpha_v > 0.$$

Generally speaking, the above optimization problem is equivalent to the binary code learning sub-problem of Eq. (11). Naturally, we can use the discrete optimization algorithm to obtain the binary code, so the updated rule of the binary code is as follows

$$\mathbf{b}_j = sgn(\mathbf{p}_j - \mathbf{h}_{vj} \mathbf{H}_v'^T \mathbf{B}'^T) \tag{36}$$

$$s.t. \mathbf{B} \in \{-1, 1\}^{k \times n}, \sum_{v=1}^{n_t} \alpha_v = 1, \alpha_v > 0.$$

where $\mathbf{P} = \sum_{v=1}^{n_t} (\alpha_v)^\gamma \mathbf{X}_v^T \mathbf{D}^T \mathbf{U}_v, \mathbf{H}_v = \overline{\mathbf{D}} \mathbf{U}_v, \overline{\mathbf{D}} = \mathbf{D}^{\frac{1}{2}}, \mathbf{D} = 1/2 \sum_{v=1}^{n_t} (\alpha_v)^\gamma \|\mathbf{X}_v - \mathbf{U}_v \mathbf{B}\|_2$, $\mathbf{b}_j$, $\mathbf{p}_j$ and $\mathbf{h}_{vj}$ stand for the jth row of **B**, **P** and $\mathbf{H}_v$, respectively. Also $\mathbf{B}'$ be the matrix of **B** excluding $\mathbf{b}_j$, $\mathbf{P}'$ be the matrix of **P** excluding $\mathbf{p}_j$, $\mathbf{H}_v'$ be the matrix of $\mathbf{H}_v$ excluding $\mathbf{h}_{vj}$, respectively. α_v is the optimal weight obtained by Algorithm 1.

4. Experiments

In this section, we apply the proposed method RMDH to three benchmark datasets to prove the effectiveness of the method. Firstly, we introduce the details of the three datasets, the evaluation criteria, the proposed comparison method, and the parameter settings in the experiment. Next, we give a comparison of the experimental results and the discussion. Finally, the problems of convergence, computational efficiency and parameter sensitivity of RMDH are further researched.

4.1. Datasets

The validity and efficiency of the proposed RMDH model are verified by using three benchmark multimodal transport datasets. In particular, we describe the statistics of these datasets and a brief description of each dataset as follows:

Wiki consists of 2866 image–text pairs, which collected from Wikipedia [51] and divided into 10 categories. Moreover, it is single-label dataset for each sample in one of them. For each sample, a 128-dimension SIFT histogram topics vector is extracted used as image feature and each text feature is represented by a 10-dimension topics vector. In our experiments, 75% of the samples were used as the training set and the rest were used as the query set.

MIRFlickr consists of 25,000 text–image pairwise data points, which collected from the Flickr website [52]. Each point is annotated by some textual tags selected from 24 labels. For each pair of images, the image is described by a 150 dimensional bag-of-words vector, and points from the text is described by a 500 dimensional eigenvector derived from the PCA. As in Ref. [53], we only keep the samples with text labels appearing at least 20 times, and delete image label pairs without text labels or manual annotation labels. Then we can get 20015 pairs of image tags for experiment. We randomly select 5% of image tag pairs as query sets and the remaining as training sets.

NUS-WIDE: this dataset consists of 269,648 pairwise data points [54], and each sample contains an image and its associated text label. We only keep the 10 most common concepts and discard samples that are not included in these concepts. We screened 186,577 corresponding examples, where these sample images were represented as 500-dimensional visual bag word SIFT features, and the text was represented as 1000-dimensional marked appearance feature vectors. We randomly select 2000 image label pairs as the query set for the experiment and the remaining as the training set.

4.2. Evaluation criteria

In order to fairly evaluate the superiority of the method, we use three common indicators mAP, top-N accuracy and precision–recall to measure and evaluate retrieval performance. The indicators are defined as follows:

(1) mAP: We use average mean accuracy (mAP) to evaluate the performance of cross-modal search results:

$$AP = \frac{1}{N} \sum_{r=1}^{R} P(r)\delta(r),$$

where N is the number of queries, $P(r)$ denotes the precision of the top r retrieved instances, $\delta(r)$ is defined as the indicator function. If the rth entity is related to the query, its value is 1 or 0 otherwise. Obviously, the larger the mAP, the better the performance.

(2) top-N precision: It reflects the change in the ratio of the number of related documents retrieved to the total number of documents retrieved.

(3) precision–recall: Precision and recall rate are mutually influential. Ideally, both are high, but in general, the precision is high, the recall rate is low, vice versa.

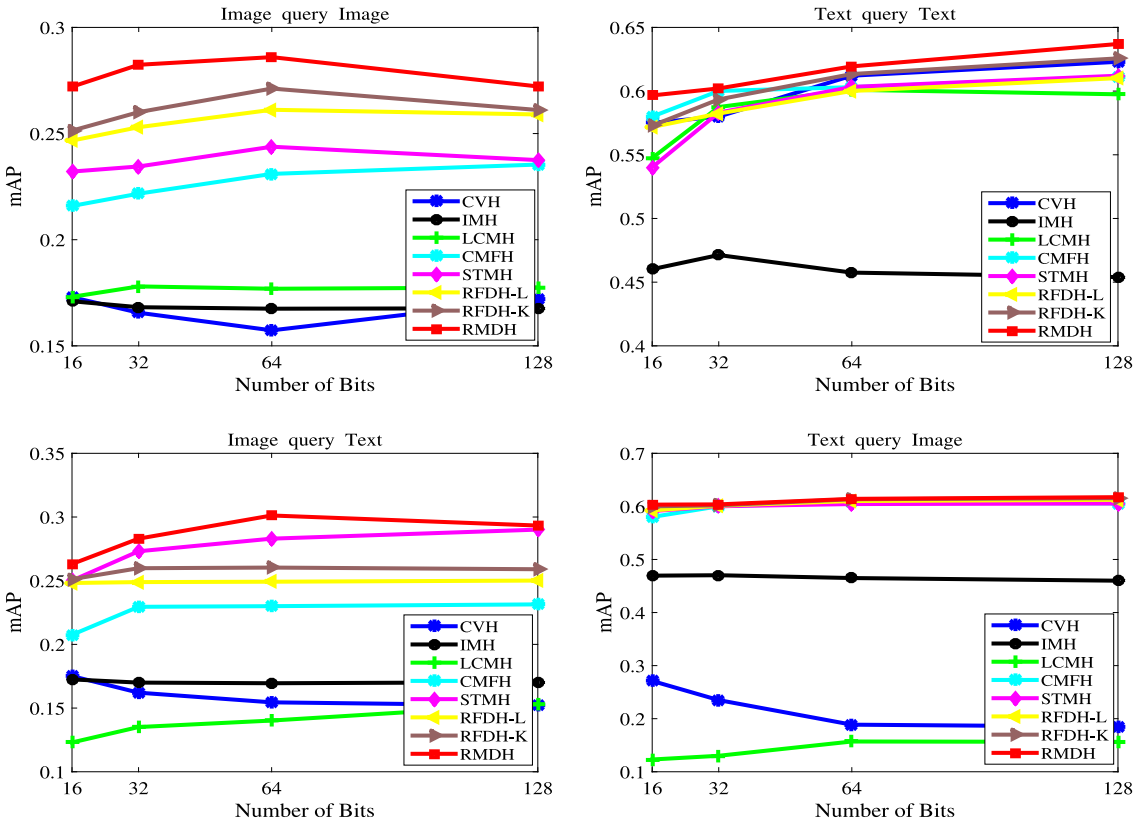

Fig. 1. mAP on Wiki by varying code length.

4.3. Baseline methods

We compare RMDH method with a series of state-of-the-art multi-modal hashing methods using the evaluation metrics described above. In order to better demonstrate the effectiveness of the proposed method, the following methods are divided into supervised and unsupervised methods in terms of the comparison of average accuracy.

- CVH [15] extends a single-modal search to a multi-modal search and use the solution to the generalized eigenvalue problem to learn the hash function by minimizing the weighted Hamming distance between the hash codes of the training data.
- IMH [37] maps the heterogeneous data to the common Hamming space by the linear hash functions, and optimizes the generation of hash code.
- LCMH [39] first calculates the distance between each training data point and the distance cluster center, and then mapping the representations to a common Hamming space to learns the hash function.
- CMFH [46] implements dimensionality reduction through collective matrix decomposition, and builds models for potential factors of different modes for learning a unified and efficient hash code.
- STMH [41] models text as multiple semantic topics, models image as a latent semantic concept, and generates a hash code by detecting whether there are concepts in the dataset.
- RFDH [53] introduces two methods of learning hash functions, projecting abstract instances into hash codes and uses linear classifier (RFDH-L) and kernel logistic (RFDH-K) regression for out-of-sample extension.

4.4. Implementation details

IMH consumes more time in calculation, and it is difficult to learn hash functions with large datasets. Therefore, we use the sampling method to randomly select 10,000 instances in the big dataset to train the hash function, then apply the trained hash function to test instances in the dataset to generate a hash code.

We perform a sensitivity analysis on the parameters, and after sufficient experiments, we verify that the RMDH can achieve stable performance in various ranges. When compare to the baseline method, we set parameter as following, i.e. $\gamma = 3, \theta = 0.001, \beta = 100$ and $\eta = 1$.

In addition to the baseline algorithms other than CVH, the authors provide these codes amicably. Since the initial CVH implementation has not made public, we implemented it ourselves. We adjust the parameters of all baseline algorithms based on the research experience and the appropriate adjustment of the parameters. In the following experiments, for IMH, CMFH, STMH and RFDH, the convergence threshold and the maximum numbers of iterations are set from 0.01 to 20, respectively.

4.5. Experimental results

(1) mAP results: For convenience, we divide all benchmark methods into supervised and unsupervised categories and study the experimental results of mAP on the Wiki, MIRFlickr and NUS-WIDE datasets, The mAP values of RMDH and six baseline methods on the dataset are recorded in Iand plotted in Figs. 1–3. From I, as the length of the hash code increases, we get no matter what kind of retrieval task is, the retrieval performance will be improved to varying degrees. For example, in the task of image query image for NUS-WIDE, the retrieval performance of STMH improves by nearly 10% for the length of hash code from 8-bits to 128-bits. On wiki, the mAP of LCMH is $0.1231, 0.1301$ and 0.1570 when the code length is 16, 32 and 64, respectively; in comparison, the MAP of the proposed RMDH is $0.6033, 0.6038$ and 0.6142 when the code length is 16, 32 and 64, respectively. These data show that even when the binary code length is short, RMDH can still show better performance. From Fig. 1, we have the following several observations. Firstly, in most cases, RMDH performs the best

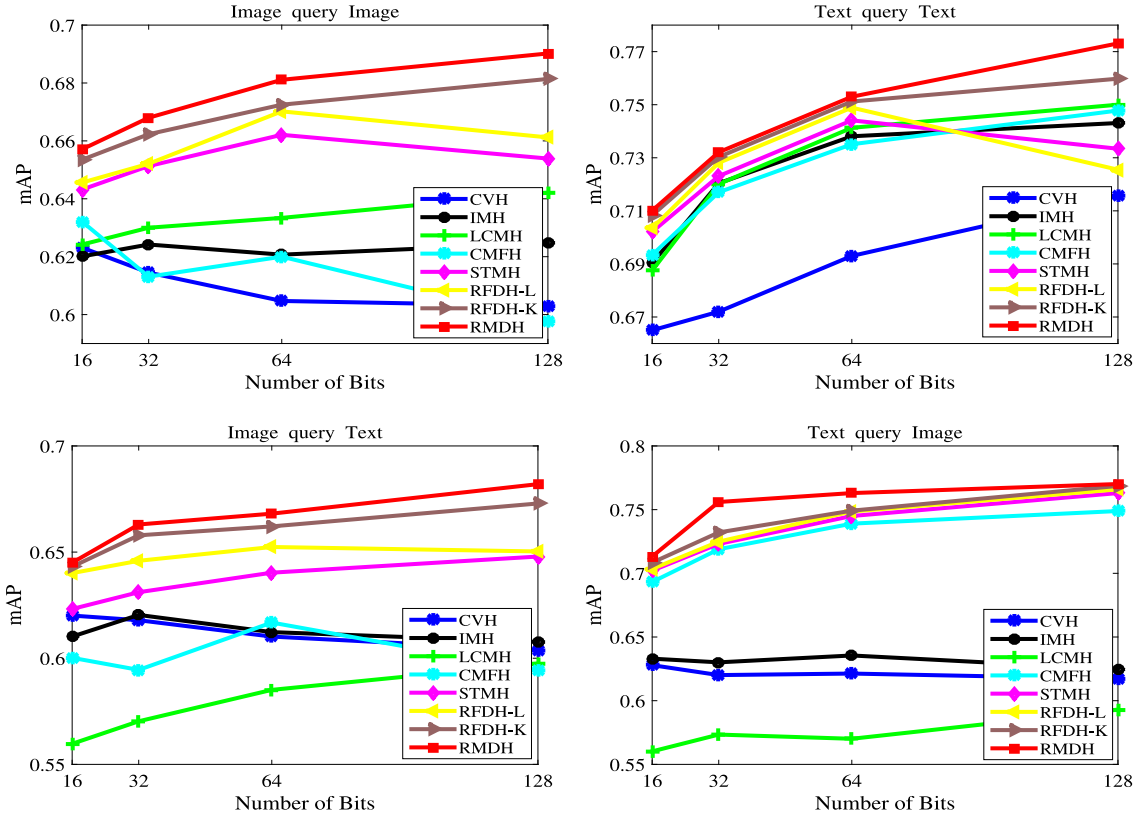

Fig. 2. mAP on MIRFlickr by varying code length.

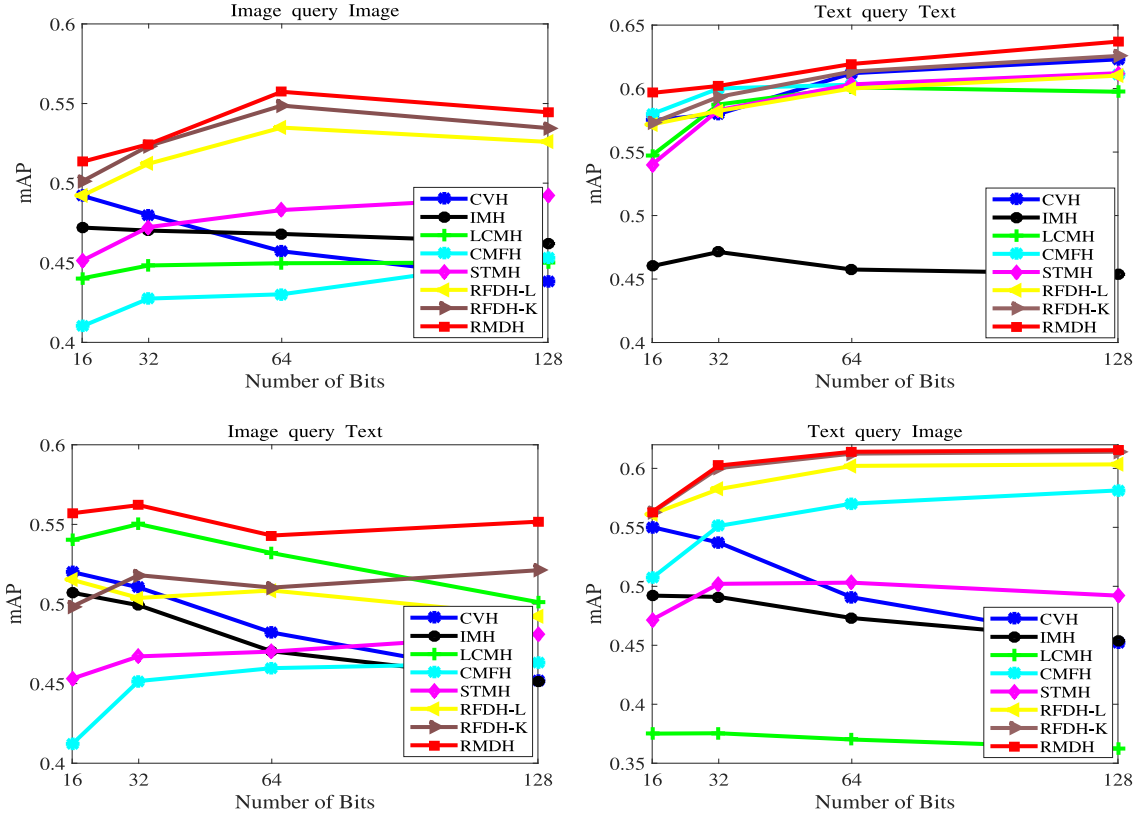

Fig. 3. mAP on NUS-WIDE by varying code length.

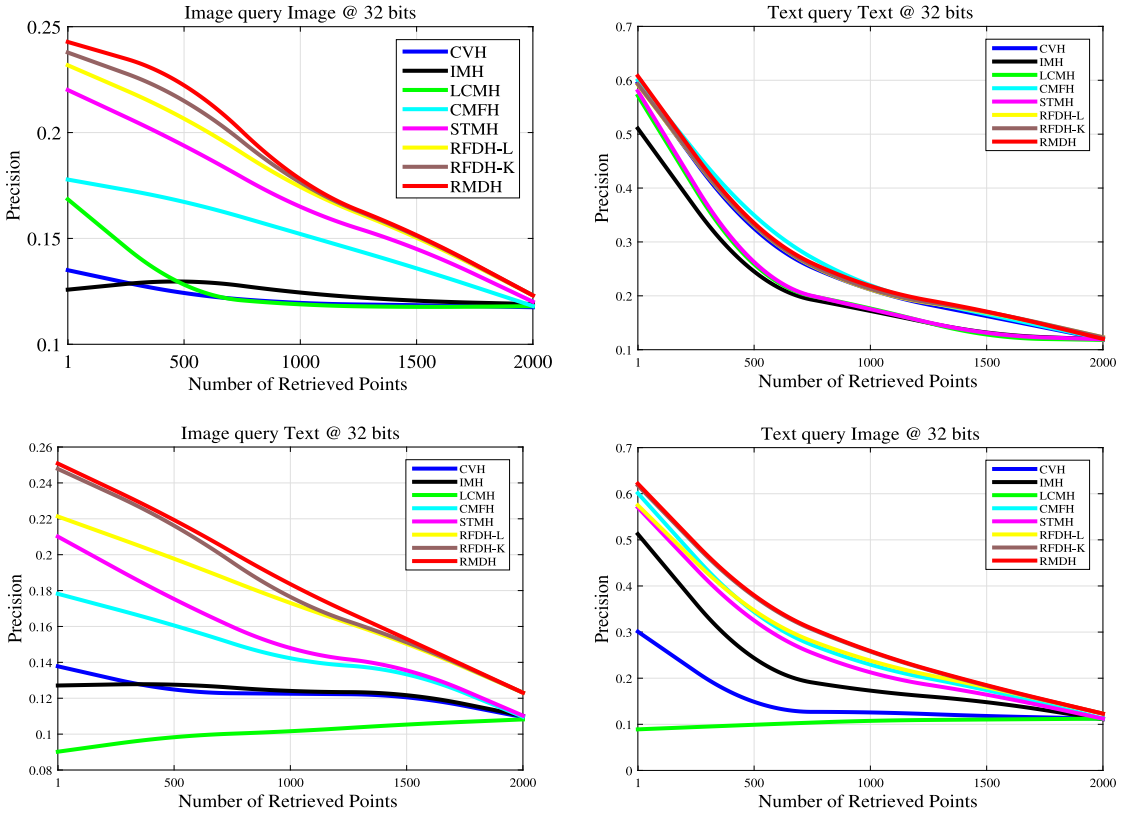

Fig. 4. top-N precision curves on Wiki by varying code length.

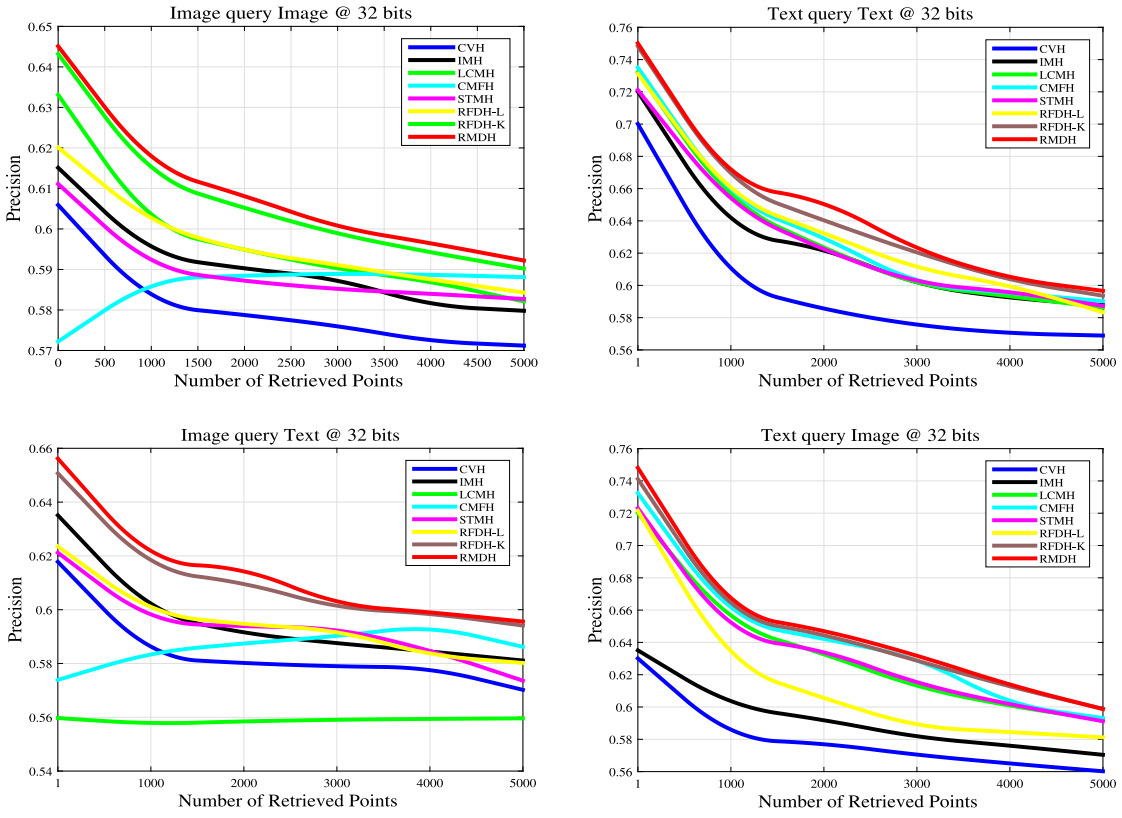

Fig. 5. top-N precision curves on MIRFlickr by varying code length.

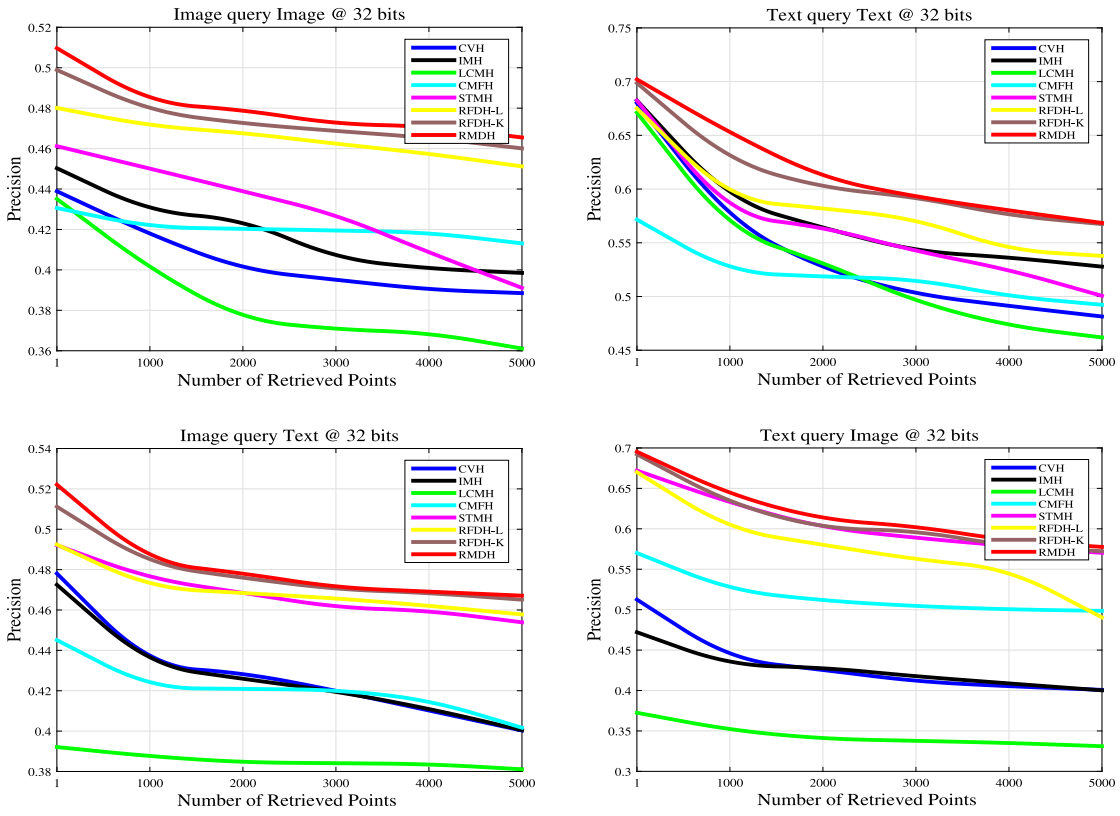

Fig. 6. top-N precision curves on NUS-WIDE by varying code length.

compared with other methods. Secondly, in the image retrieval image, the mAP value of RMDH is slightly lower than the mAP value of STMH, but is far superior to the others. Thirdly, the mAP value of the text query in all methods is better than the mAP value of the image query. On the other hand, the extracted sample attributes do not reflect the semantic properties very well, resulting in the waste of resources in the calculation (see Table 1).

(2) top-N precision: The top-N precision curves are plotted in Figs. 4–6, respectively. We have three observations from the experimental results. Consistent with mAP results, it is easy to get the top-N of RMDH from the Figs. 4–6, which is better than all the comparison methods. Secondly, the experiment result shows, for all methods, the change of top-N Precision curves is fast between 500 and 1000 search points. Thirdly, RMDH shows the storage efficiency and retrieval efficiency of our method. For example, in some cases, even relative short binary codes (16 bits), RMDH still defeats the compared methods compare with longer binary codes (32 bits).

(3) precision–recall: The precision–recall curves with the code length of 32 bits is also shown in Figs. 7–9. By observing the position relationship of the precision–recall curves, it is clear that the proposed RMDH always provides better performance than every other prior art methods. It is worth mentioning that when the recall rate increases, the accuracy of the proposed method decreases less. We can also observe RMDH demonstrates good robustness in completing cross-media retrieval tasks. On dataset MIRFlickr, RFDH-L and RFDH-K always outstanding other methods, RFDH-K reached the suboptimal accuracy due to it used nonlinear classifier and learn binary codes by matrix decomposition, avoiding the quantization errors caused by relaxation.

4.6. Parameter sensitivity

In this section, we analyze the effects of different parameter settings on the performance of the algorithm through theoretical analysis and experiments. There are four parameters in our RMDH, including γ, β and η in the objective function. In our experimental setup, we empirically set $\gamma = 3, \beta = 100$ and $\eta = 1$. In order to better compare the influence of parameters, we fixed the hash code length of 16 bits, and study the influence of the parameters on the retrieval effect by changing one parameter and fixing the other one. We take the change in the experimental value mAP as an evaluation of the performance change with respect to different parameter values and plot Fig. 6. We know that all parameters have good robustness, and RMDH can produce satisfactory results in various parameter values.

It is worth mentioning that the magnitude of the effect on the parameter α_v largely depends on γ. For Eq. (34).

$$\lim_{\gamma\to\infty}\alpha_v = \lim_{\gamma\to\infty}\frac{(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}}{\sum_{v=1}^{2}(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}} = \frac{1}{2}.$$

$$\lim_{\gamma\to1^+}\alpha_v = \lim_{\gamma\to1^+}\frac{(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}}{\sum_{v=1}^{2}(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}} = \begin{cases} 1, \mathbf{H}_v = max\{\mathbf{H}_1, \mathbf{H}_2\}, \\ 0, \mathbf{H}_v = min\{\mathbf{H}_1, \mathbf{H}_2\}, \end{cases}$$

$$\lim_{\gamma\to1^-}\alpha_v = \lim_{\gamma\to1^-}\frac{(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}}{\sum_{v=1}^{2}(\gamma\mathbf{H}_v)^{\frac{1}{\gamma-1}}} = \begin{cases} 1, \mathbf{H}_v = min\{\mathbf{H}_1, \mathbf{H}_2\}, \\ 0, \mathbf{H}_v = max\{\mathbf{H}_1, \mathbf{H}_2\}, \end{cases}$$

This shows that when $\gamma \to \infty$, the weighting factors of different modes are equal. On the other hand, from the above analysis, we have that the weight of the modal with the smallest error $\mathbf{H}_v$ will be 1 and 0 otherwise. Therefore, the choice of γ should be based on the principle of complementarity of all modals. In other words, if each mode plays a similar role, then we should choose a larger number, otherwise, γ should be small. we get from Fig. 10 that, image query text performs the best around $\gamma = 6$ and text retrieval image performs the best around $\gamma = 3$ when the hash code length is 32 bits. Experiments show that the proposed RMDH can achieve stable performance over a wide range of γ.

Table 1
mAP results on different datasets. The best performance is shown in boldface.

Task	Category	Method	Wiki				MIRFlickr				NUS-WIDE			
			16-bit	32-bit	64-bit	128-bit	16-bit	32-bit	64-bit	128-bit	16-bit	32-bit	64-bit	128-bit
Image query Image	Supervised	CVH	0.1729	0.1656	0.1573	0.1720	0.6231	0.6145	0.6047	0.6030	0.4921	0.4800	0.4572	0.4381
		STMH	0.2321	0.2345	0.2438	0.2375	0.6432	0.6513	0.6621	0.6539	0.4513	0.4724	0.4831	0.4921
	Unsupervised	LCMH	0.1731	0.1780	0.1769	0.1775	0.6242	0.6300	0.6334	0.6421	0.4402	0.4483	0.4497	0.4502
		CMFH	0.2159	0.2217	0.2310	0.2354	0.6321	0.6131	0.6200	0.5978	0.4101	0.4275	0.4302	0.4534
		IMH	0.1710	0.1682	0.1675	0.1677	0.6201	0.6242	0.6207	0.6247	0.4721	0.4702	0.4681	0.4623
		RFDH_L	0.2467	0.2530	0.2612	0.2590	0.6456	0.6521	0.6702	0.6612	0.4921	0.5122	0.5349	0.5259
		RFDH_K	0.2514	0.2601	0.2712	0.2611	0.6535	0.6623	0.6725	0.6814	0.5011	0.5234	0.5487	0.5346
		RMDH	**0.2721**	**0.2824**	**0.2860**	**0.2722**	**0.6544**	**0.6663**	**0.6775**	**0.6854**	**0.5134**	**0.5245**	**0.5574**	**0.5445**
Text query Text	Supervised	CVH	0.5751	0.5801	0.6121	0.6230	0.6650	0.6720	0.6930	0.7157	0.6301	0.6423	0.6531	0.6572
		STMH	0.5402	0.5831	0.6033	0.6121	0.7021	0.7231	0.7441	0.7334	0.5145	0.5632	0.5772	0.5780
	Unsupervised	LCMH	0.5475	0.5875	0.6010	0.5975	0.6875	0.7201	0.7413	0.7500	0.5503	0.6270	0.6502	0.6631
		CMFH	0.5800	0.6001	0.6031	0.6112	0.6934	0.7170	0.7351	0.7478	0.5070	0.5541	0.5679	0.5715
		IMH	0.4602	0.4713	0.4575	0.4540	0.6903	0.7204	0.7381	0.7431	0.5724	0.6211	0.6321	0.6473
		RFDH_L	0.5718	0.5824	0.6001	0.6102	0.7035	0.7281	0.7489	0.7255	0.6211	0.6304	0.6431	0.6516
		RFDH_K	0.5729	0.5935	0.6134	0.6257	0.7083	0.7303	0.7512	0.7598	0.6214	0.6459	0.6584	0.6652
		RMDH	**0.5966**	**0.6021**	**0.6194**	**0.6370**	**0.7101**	**0.7321**	**0.7530**	**0.7731**	**0.6234**	**0.6543**	**0.6614**	**0.6690**
Image query Text	Supervised	CVH	0.1751	0.1621	0.1545	0.1520	0.6201	0.6179	0.6102	0.6038	0.5201	0.5104	0.4821	0.4521
		STMH	0.2501	0.2731	0.2830	0.2901	0.6232	0.6311	0.6403	0.6479	0.4531	0.4671	0.4701	0.4811
	Unsupervised	LCMH	0.1230	0.1351	0.1402	0.1531	0.5597	0.5703	0.5851	0.5975	0.5401	0.5502	0.5320	0.5012
		CMFH	0.2075	0.2295	0.2300	0.2314	0.6002	0.5945	0.6168	0.5943	0.4120	0.4516	0.4597	0.4631
		IMH	0.1725	0.1700	0.1695	0.1703	0.6102	0.6204	0.6123	0.6077	0.5071	0.4992	0.4703	0.4511
		RFDH_L	0.2482	0.2489	0.2492	0.2501	0.6401	0.6459	0.6524	0.6503	0.5152	0.5038	0.5085	0.4921
		RFDH_K	0.2512	0.2598	0.2603	0.2590	0.6431	0.6579	0.6621	0.6729	0.4984	0.5181	0.5103	0.5212
		RMDH	**0.2631**	**0.2830**	**0.3012**	**0.2933**	**0.6450**	**0.6630**	**0.6681**	**0.6820**	**0.5470**	**0.5621**	**0.5430**	**0.5447**
Text query Image	Supervised	CVH	0.2711	0.2350	0.1887	0.1850	0.6279	0.6201	0.6213	0.6172	0.5502	0.5371	0.4907	0.4521
		STMH	0.5911	0.6003	0.6044	0.6051	0.7021	0.7231	0.7450	0.7630	0.4717	0.5021	0.5031	0.4921
	Unsupervised	LCMH	0.1231	0.1301	0.1570	0.1563	0.5602	0.5734	0.5701	0.5930	0.3751	0.3754	0.3701	0.3624
		CMFH	0.5801	0.6002	0.6075	0.6055	0.6932	0.7189	0.7390	0.7489	0.5071	0.5512	0.5701	0.5813
		IMH	0.4695	0.4703	0.4650	0.4600	0.6330	0.6301	0.6355	0.6247	0.4921	0.4910	0.4731	0.4534
		RFDH_L	0.5924	0.6013	0.6101	0.6124	0.7034	0.7249	0.7485	0.7659	0.5612	0.5824	0.6021	0.6034
		RFDH_K	0.6001	0.6021	0.6135	0.6155	0.7084	0.7321	0.7492	0.7686	0.5625	0.6002	0.6125	0.6141
		RMDH	**0.6033**	**0.6038**	**0.6142**	**0.6174**	**0.7132**	**0.7560**	**0.7630**	**0.7701**	**0.5631**	**0.6024**	**0.6141**	**0.6154**

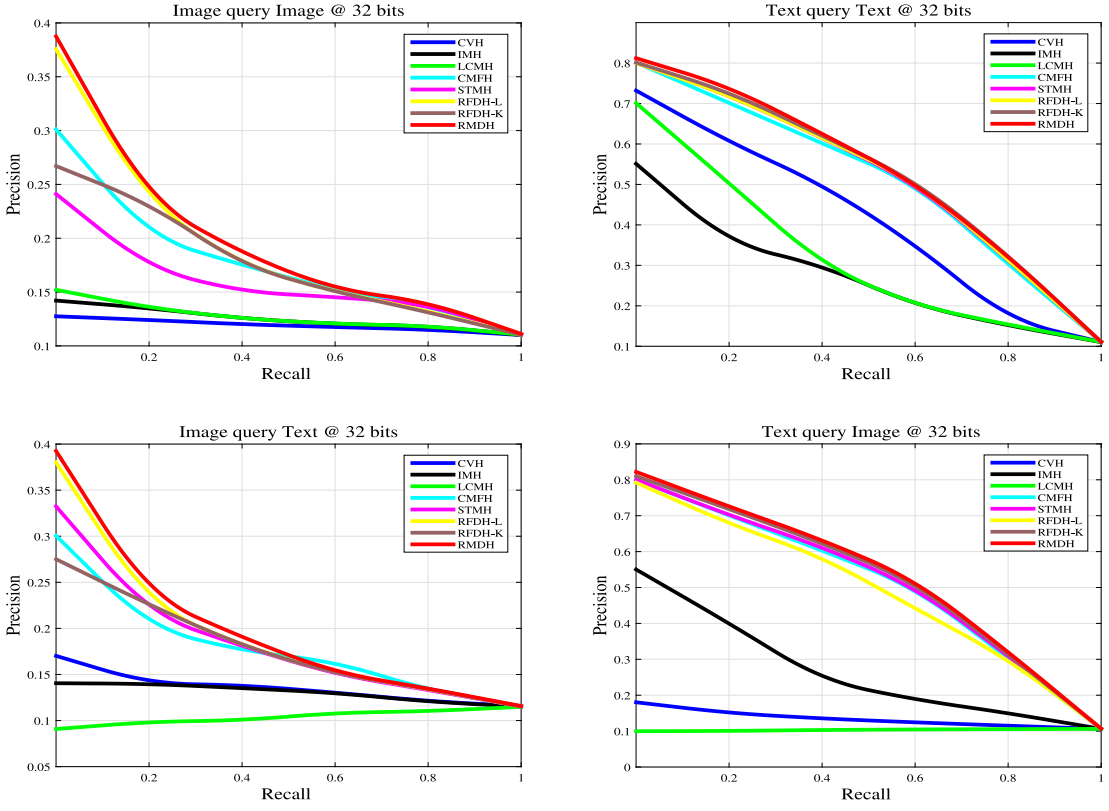

Fig. 7. Precision–recall curves on Wiki by varying code length.

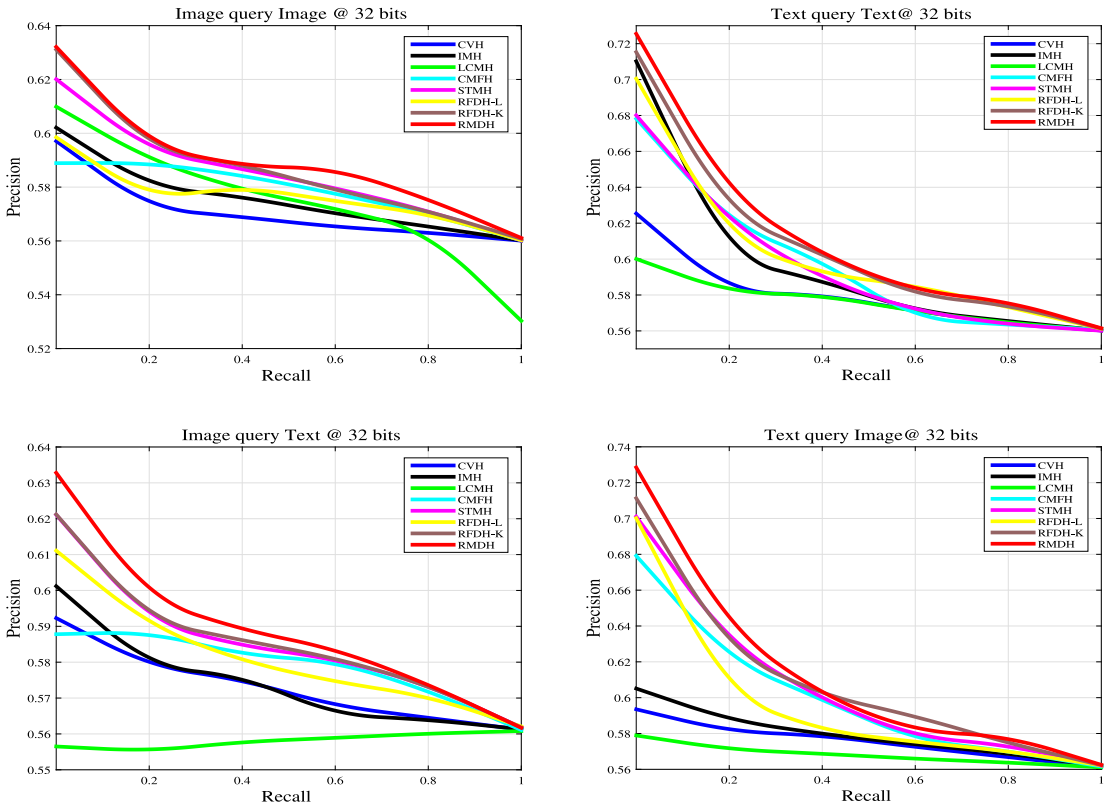

Fig. 8. Precision–recall curves on MIRFlickr by varying code length.

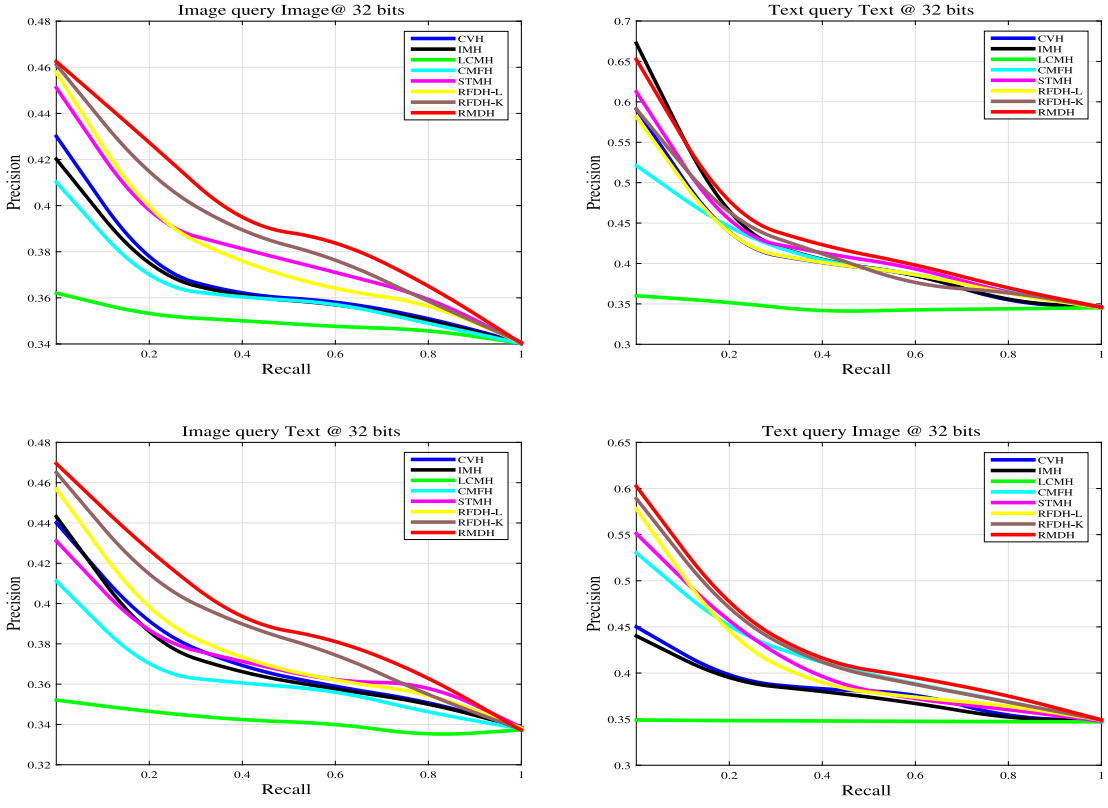

Fig. 9. Precision–recall curves on NUS-WIDE by varying code length.

β mainly controls the learning term of the hash function, which in turn affects the objective function. If its value is too small or too large, it will reduce the effectiveness of the hash function. From Fig. 10, it can be found that when performing the task of image retrieval text on the

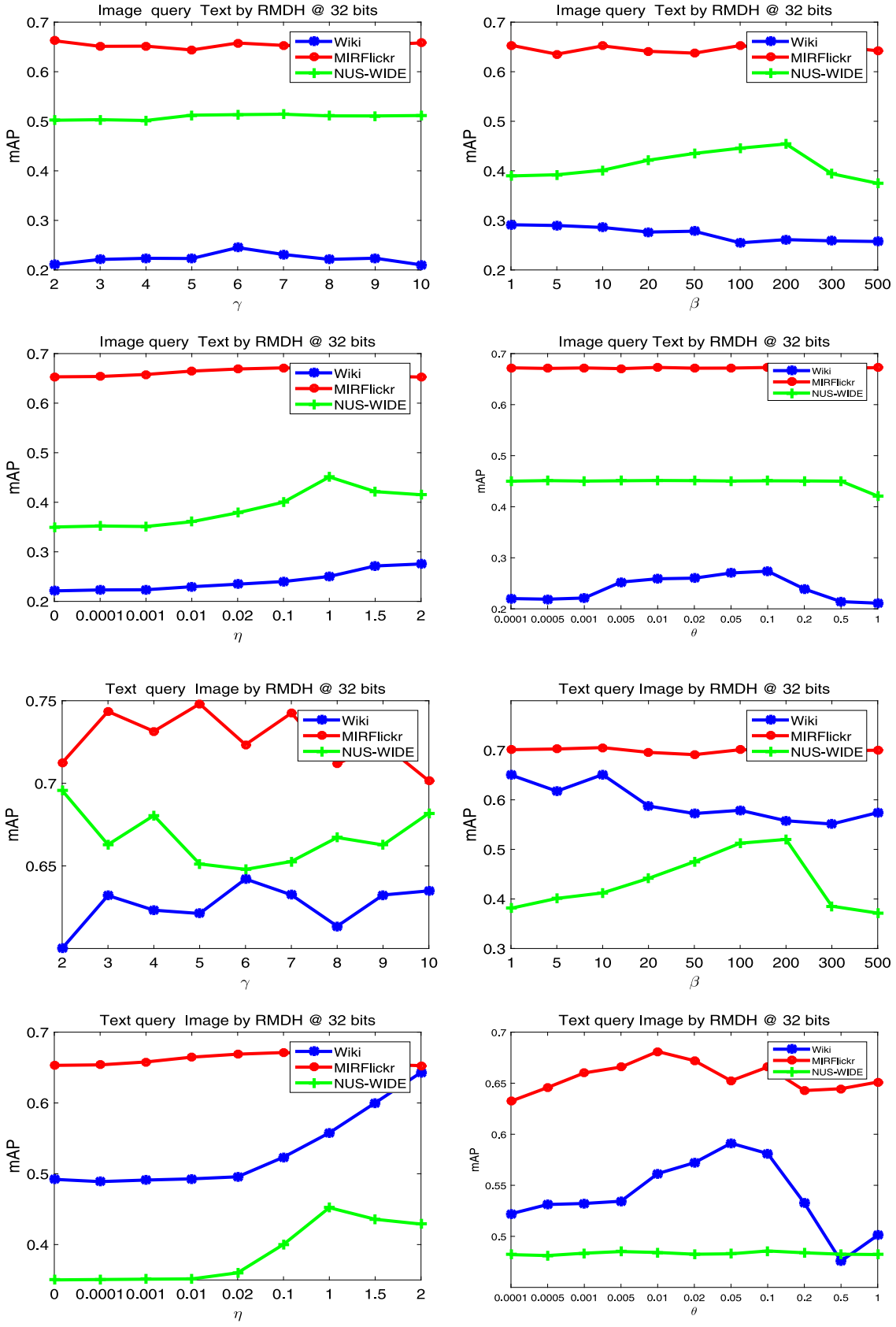

Fig. 10. Parameter sensitivity analysis.

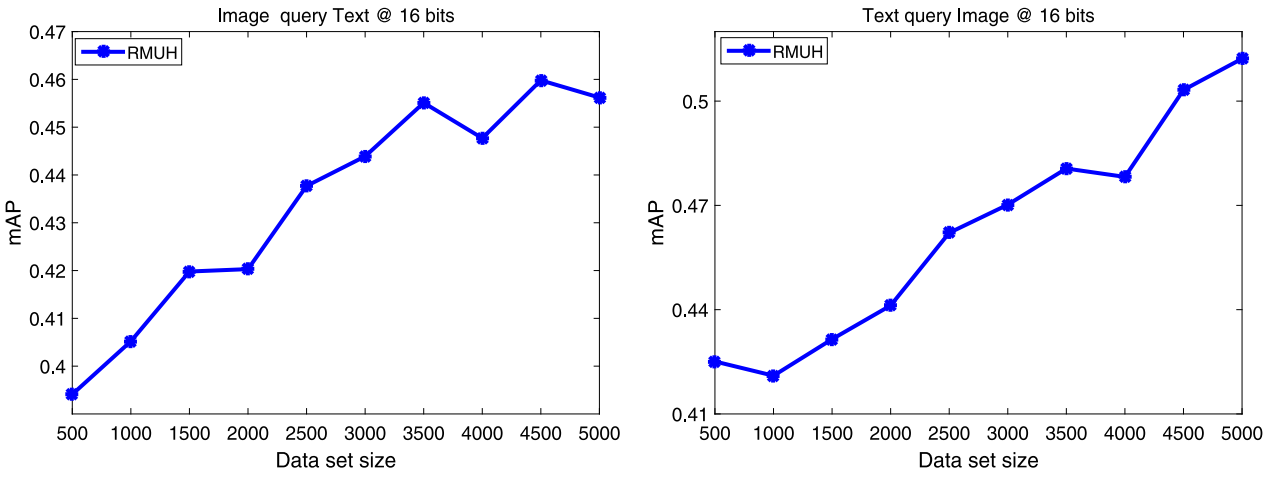

Fig. 11. Effect of dataset size on mAP on the NUS-WIDE dataset.

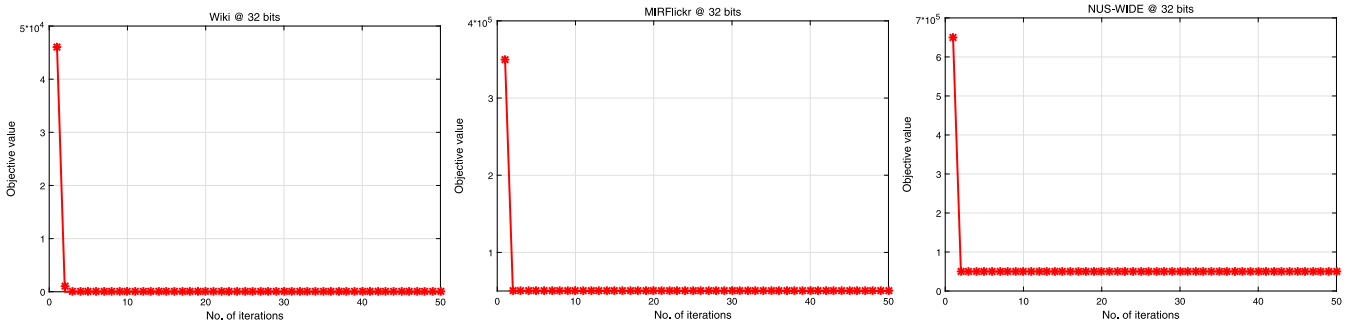

Fig. 12. Convergence curves.

Wiki dataset, with the increase of β, when the hash code length is 32 bits, the mAP value of the RMUH first increases, and then decreases until it stabilizes. Similar situation also occurs in the case of text retrieval images. In order to better achieve the retrieval effect, it is not difficult to select $\beta \in [10, 200]$.

θ is a parameter introduced when updating parameter $\mathbf{U}_v$, in order to prevent the occurrence of over-fitting. Usually, θ can be chosen from the range between [0.001, 0.2].

The parameter η holds the semantic similarity of the original space, which is set as a large number of 1. We can see that the experiment results are not very sensitive to parameter η.

From the above analysis, it can be obtained that the proposed model has good stability to the parameters. On the whole, these parameters can be selected in a larger range and have good scalability.

4.7. Effect analysis of $\ell_{2,1}$-norm

The introduction of the $\ell_{2,1}$-norm as the loss function is one of the main contributions of the proposed RMDH method. In order to better verify the advantages of the loss function in mitigating the negative impact of noisy data and enhancing the stability of the model, we compare it with the well-known ℓ_2-norm on the dataset NUS-WIDE for retrieval performance. For convenience, we denote RMDH-2 to represent the model using the ℓ_2-norm.

Fig. 13 shows the mAP values of RMDH and RMDH-2 models with different hash code lengths on the NUS-WIDE dataset. It is easy to see that RMDH will achieve better performance than the ℓ_2-norm when using the $\ell_{2,1}$-norm.

4.8. Data set size

In this section, we analyze the effects of different dataset sizes on NUS-WIDE datasets. For convenience, we fix the length of hash code to 16 bits and study it by changing the size of dataset. We set the dataset from 500 to 5000, and increase the data number by 500 each time. It is easy to see that the value of mAP increases with the increase of datasets. Experiments show that RMDH is not only suitable for large-scale datasets, but also can achieve satisfactory performance on smaller datasets (see Fig. 11).

4.9. Convergence analysis

As the Eq. (6) shows, there are six different variables. It is worth noting that we propose an iterative optimization scheme to solve this non-convex optimization problem. Therefore, we conduct experiments on three datasets to confirm the convergence of the optimization algorithm, Fig. 12 shows the convergence curve at 32 bits. On these three datasets, as shown in Fig. 12, RMDH can converge within 10 iterations.

5. Conclusion

In this paper, we have proposed an effective and robust cross-modal hashing approach named Robust Multimodal Discrete Hashing for Cross-modal Similarity Search. Firstly, the proposed approach uses collective matrix factorization to generate unified hash codes to achieve cross-modal search. Secondly, by directly learning discrete hash codes, the hash modal avoids large quantization errors and adopts $\ell_{2,1}$ norm which is more robust to disturbing data. In addition, it adaptively adjusts the weights of each modal in the process of hash code learning. We introduce a mixed graph of Laplacian regularization terms to improve

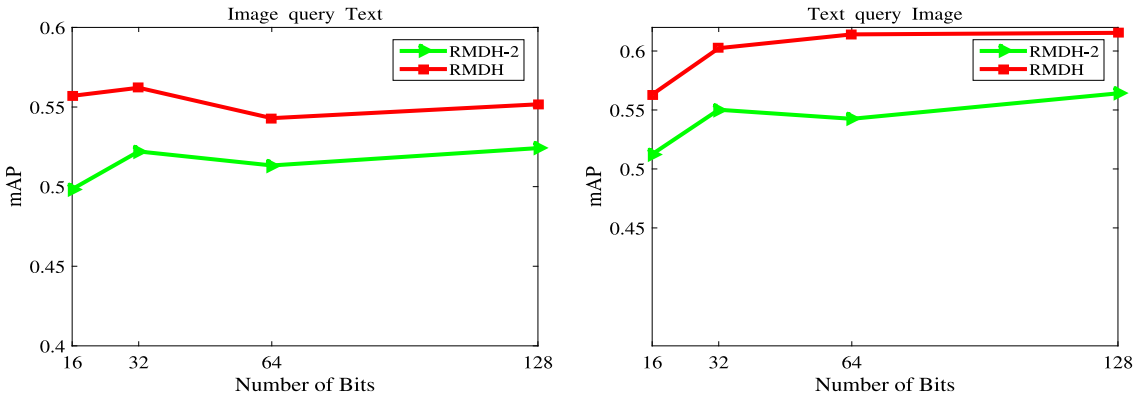

Fig. 13. mAP values of the RMDH variants on NUS-WIDE.

the retrieval efficiency and efficiency. Extensive experiments on three benchmark datasets showed the effectiveness of the proposed RMDH, as compared several state-of-the-art methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to reviewers and editors for their hard work.

References

[1] J. Wang, T. Zhang, J. Song, N. Sebe, H.T. Shen, A survey on learning to hash, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2016) 769–790.
[2] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
[3] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, Neural Comput. 12 (6) (2000) 1247–1283.
[4] N. Gao, S.-J. Huang, Y. Yan, S. Chen, Cross modal similarity learning with active queries, Pattern Recognit. 75 (2018) 214–222.
[5] D. Ngo, A. Teoh, A. Goh, Biometric hash: high-confidence face recognition, IEEE Trans. Circuits Syst. Video Technol. 16 (6) (2006) 771–775.
[6] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2012) 2916–2929.
[7] C. Xiao, F. Nie, W. Cai, H. Huang, Heterogeneous image features integration via multi-modal semi-supervised learning model, in: 2013 IEEE International Conference on Computer Vision, 2014, pp. 1737–1744.
[8] X. Liu, A. Li, J.-X. Du, S.-J. Peng, W. Fan, Efficient cross-modal retrieval via flexible supervised collective matrix factorization hashing, Multimedia Tools Appl. 77 (21) (2018) 28665–28683.
[9] G. Lin, C. Shen, D. Suter, A. Van Den Hengel, A general two-step approach to learning-based hashing, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2552–2559.
[10] X. Dong, J. Sun, P. Duan, L. Meng, Y. Tan, W. Wan, H. Wu, B. Zhang, H. Zhang, Semi-supervised modality-dependent cross-media retrieval, Multimedia Tools Appl. 77 (3) (2018) 3579–3595.
[11] Y. Peng, X. Zhai, Y. Zhao, X. Huang, Semi-supervised cross-media feature learning with unified patch graph regularization, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 583–596.
[12] L. Zhu, Z. Huang, Z. Li, L. Xie, H.T. Shen, Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval, IEEE Trans. Neural Netw. Learn. Syst. 29 (11) (2018) 5264–5276.
[13] D.H. Greene, M. Parnas, F.F. Yao, Multi-index hashing for information retrieval, in: 35th Annual Symposium on Foundations of Computer Science, Santa Fe, New Mexico, USA, 20-22 November 1994, 1994, pp. 722–731.
[14] X. Lu, L. Zhu, Z. Cheng, X. Song, H. Zhang, Efficient discrete latent semantic hashing for scalable cross-modal retrieval, Signal Process. 154 (2019) 217–231.
[15] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, 2011.
[16] Z. Lin, G. Ding, M. Hu, J. Wang, Semantics-preserving hashing for cross-view retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3864–3872.
[17] W. Liu, C. Mu, S. Kumar, S.-F. Chang, Discrete graph hashing, Adv. Neural Inf. Process. Syst. 4 (2014) 3419–3427.
[18] Y. Yang, F. Shen, H.T. Shen, H. Li, X. Li, Robust discrete spectral hashing for large-scale image semantic indexing, IEEE Trans. Big Data 1 (4) (2016) 162–171.
[19] M.M. Bronstein, A.M. Bronstein, F. Michel, N. Paragios, Data fusion through cross-modality metric learning using similarity-sensitive hashing, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, 2010, pp. 3594–3601.
[20] Y. Zhen, D.-Y. Yeung, A probabilistic model for multimodal hash function learning, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 940–948.
[21] D. Zhang, W.-J. Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2177–2183.
[22] D. Wang, X. Gao, X. Wang, L. He, B. Yuan, Multimodal discriminative binary embedding for large-scale cross-modal retrieval, IEEE Trans. Image Process. 25 (10) (2016) 4540–4554.
[23] J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3460–3469.
[24] S. hua Zhong, Y. Liu, Y. Liu, Bilinear deep learning for image classification, in: Proceedings of the 19th International Conference on Multimedea 2011, Scottsdale, AZ, USA, November 28 - December 1, 2011, 2011, pp. 343–352.
[25] E. Flores, M. Zortea, J. Scharcanski, Dictionaries of deep features for land-use scene classification of very high spatial resolution images, Pattern Recognit. 89 (2019) 32–44.
[26] H. Guo, K. Zheng, X. Fan, H. Yu, S. Wang, Visual attention consistency under image transforms for multi-label image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, 2019, pp. 729–739.
[27] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., Deep speech: Scaling up end-to-end speech recognition, 2014, arXiv preprint arXiv:1412.5567.
[28] Y. Guo, T. Chen, Semantic segmentation of rgbd images based on deep depth regression, Pattern Recognit. Lett. 109 (2018) 55–64.
[29] X. Zhou, W. Gong, W. Fu, F. Du, Application of deep learning in object detection, in: 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 2017, pp. 631–634.
[30] R. Hou, M. Pan, Y. Zhao, Y. Yang, Image anomaly detection for iot equipment based on deep learning, J. Vis. Commun. Image Represent. 64 (2019) 102599.
[31] K. Fu, Q. Zhao, I.Y.-H. Gu, J. Yang, Deepside: A general deep framework for salient object detection, Neurocomputing 356 (2019) 69–82.
[32] L. Liu, M. Yu, L. Shao, Latent structure preserving hashing, Int. J. Comput. Vis. 122 (3) (2017) 439–457.
[33] Q. Ma, C. Bai, J. Zhang, Z. Liu, S. Chen, Supervised learning based discrete hashing for image retrieval, Pattern Recognit. 92 (2019) 156–164.
[34] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Trans. Image Process. 27 (8) (2018) 3893–3903.
[35] C. Wang, H. Yang, C. Meinel, Deep semantic mapping for cross-modal retrieval, in: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2015, pp. 234–241.
[36] Y. Shen, L. Liu, L. Shao, J. Song, Deep binaries: Encoding semantic-rich cues for efficient textual-visual cross retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4097–4106.

[37] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, ACM, 2013, pp. 785–796.

[38] X. Zhu, Z. Huang, H.T. Shen, X. Zhao, Linear cross-modal hashing for efficient multimedia search, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 143–152.

[39] D. Wang, X.-B. Gao, X. Wang, L. He, Label consistent matrix factorization hashing for large-scale cross-modal similarity search, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 2466–2479.

[40] D.C.G. aes Pedronette, F.M.F. Gonçalves, I.R. Guilherme, Unsupervised manifold learning through reciprocal knn graph and connected components for image retrieval tasks, Pattern Recognit. 75 (2018) 161–174.

[41] D. Wang, X. Gao, X. Wang, L. He, Semantic topic multimodal hashing for cross-media retrieval, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 3890–3896.

[42] L. He, X. Xu, H. Lu, Y. Yang, H.T. Shen, Unsupervised cross-modal retrieval through adversarial learning, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 1153–1158.

[43] D. Wang, Q. Wang, Y. An, X. Gao, Y. Tian, Online collective matrix factorization hashing for large-scale cross-media retrieval, in: SIGIR '20: The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1409–1418.

[44] X. Lu, L. Zhu, Z. Cheng, J. Li, X. Nie, H. Zhang, Flexible online multi-modal hashing for large-scale multimedia retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, 2019, pp. 1129–1137.

[45] A.P. Singh, G.J. Gordon, Relational learning via collective matrix factorization, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 650–658.

[46] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: Computer Vision and Pattern Recognition, 2014, pp. 2083–2090.

[47] H. Liu, R. Ji, Y. Wu, G. Hua, Supervised matrix factorization for cross-modality hashing, in: International Joint Conference on Artificial Intelligence, 2016, pp. 1767–1773.

[48] J. Tang, K. Wang, L. Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Trans. Image Process. 25 (7) (2016) 3157–3166.

[49] F. Shen, C. Shen, W. Liu, H. Tao Shen, Supervised discrete hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 37–45.

[50] S.-G. Lee, Q.-P. Vu, Simultaneous solutions of sylvester equations and idempotent matrices separating the joint spectrum, Linear Algebra Appl. 435 (9) (2011) 2097–2109.

[51] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[52] M.J. Huiskes, M.S. Lew, The mir flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, ACM, 2008, pp. 39–43.

[53] D. Wang, Q. Wang, X. Gao, Robust and flexible discrete hashing for cross-modal similarity search, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2703–2715.

[54] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, 2009, p. 48.