Journal of Ambient Intelligence and Humanized Computing (2022) 13:5067–5081
https://doi.org/10.1007/s12652-020-02177-7

ORIGINAL RESEARCH

Modality-specific matrix factorization hashing for cross-modal retrieval

Haixia Xiong[1] · Weihua Ou[1,2,3] · Zengxian Yan[3] · Jianping Gou[4] · Quan Zhou[5] · Anzhi Wang[1]

Received: 10 April 2020 / Accepted: 4 June 2020 / Published online: 19 June 2020

Abstract

Cross-modal retrieval has been attracted attentively in the past years. Recently, the collective matrix factorization was proposed to learn the common representations for cross-modal retrieval based on assumption that the pairwise data from different modalities should have the same common semantic representations. However, this unified common representation could inherently sacrifice the modality-specific representations for each modality because the distributions and representations of different modalities are inconsistent. To mitigate this problem, in this paper, we propose Modality-specific Matrix Factorization Hashing (MsMFH) via alignment, which learns the modality-specific semantic representation for each modality and then aligns the representations via the correlation information. Specifically, we factorize the original feature representations into individual latent semantic representations, and then align the distributions of individual latent semantic representations via an orthogonal transformation. Then, we embed the class label into the hash codes learning via latent semantic space, and obtain hash codes directly by an efficient optimization with a closed solution. Extensive experimental results on three public datasets demonstrate that the proposed method outperforms to many existing cross-modal hashing methods up to 3% in term of mean average precision (mAP).

Keywords Cross-modal retrieval · Matrix factorization · Common semantic representations · Modality-specific · Alignment

1 Introduction

With explosive growth of multimedia data in social networks, cross-modal retrieval has attracted lots of attentions in recent years (Xu 2017; Peng et al. 2017; Deng et al. 2018; Lu et al. 2018; Yaotao et al. 2019). The task of cross-modal retrieval is to submit a query from one modality and return results in term of other different modalities (Lichao et al. 2018; Ou et al. 2019; Deng et al. 2019; Xu et al. 2019). Compared to single-modal retrieval methods (Bibi et al. 2020; Li and Zhou 2020; Hussain and Surendran 2020), cross-modal retrieval can return more comprehensive results, which contains information from different modalities. However, the fact that heterogeneous property of representations and distribution, the semantic gap between the representation and semantic label, makes cross-modal retrieval to be more challenging.

To address those issues, many approaches (Likai Qi and Hua 2018; Yang et al. 2008) have been proposed. The main principle is to construct a common semantic space utilizing the correlation between different modalities, and in which the similarities between different modalities can be measured through certain metrics. The traditional methods learns the common space by maximizing the correlations between pairwise data from different modalities, such as methods based on canonical correlation analysis (CCA) (Rasiwasia et al. 2010), kernel-CCA (Akaho 2006; Gong et al. 2014;

✉ Weihua Ou
ouweihuahust@gmail.com

[1] School of Mathematics and Sciences, School of Big Data and Computer Science, Guizhou Normal University, Guiyang, People's Republic of China

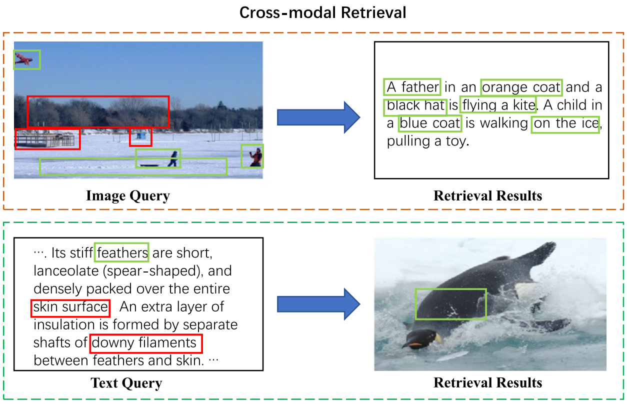
[2] Special Key Laboratory of Artificial Intelligence and Intelligent Control of Guizhou Province, Guiyang, People's Republic of China

[3] Department of Information and Electrical Engineering, Guangxi Modern Polytechnic College, Hechi, People's Republic of China

[4] School of Computer Science and Telecommunication Engineering, Jiangsu University, Jiangsu, People's Republic of China

[5] National Engineering Research Center of Communications and Networking, Nanjing University of Posts & Telecommunications, Nanjing, People's Republic of China

Wang and Livescu 2015), and deep CCA (Andrew et al. 2013). Due to the low storage and fast retrieval speed, hashing has been introduced to cross-modal retrieval (Ding et al. 2014; Chen et al. 2019; Huimin et al. 2020). The task of cross-modal hashing is to transform original feature space into hashing space, and then conducted retrieval by XOR operations efficiently. The problems of hashing are the information loss and how to preserve the correlations between different modalities during the transformation.



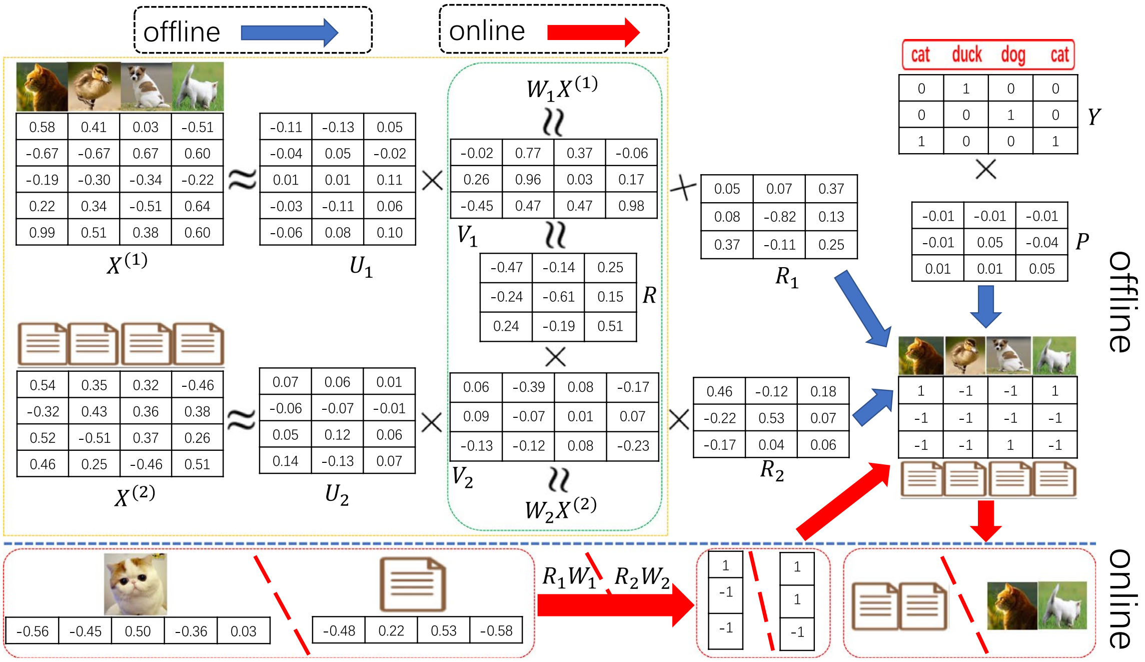
A father in an orange coat and a black hat is flying a kite. A child in a blue coat is walking on the ice, pulling a toy.

···. Its stiff feathers are short, lanceolate (spear-shaped), and densely packed over the entire skin surface. An extra layer of insulation is formed by separate shafts of downy filaments between feathers and skin. ···

Fig. 1 An example of information imbalance between different modalities, where the green boxes mean the common information described by both image and the text, while the red boxes mean the information contained in only one modality. The data is from wikepedia dataset, which is detailed in the Sect.5

Recently, collective matrix factorization hashing (Ding et al. 2014; Singh and Gordon 2008; Wang and Lu 2013; Ding et al. 2016; Li et al. 2018) have been proposed for cross-modal retrieval, based on the assumption that the pairwise data should have the same latent semantic representation (Xu et al. 2017; Jacobs et al. 2012; Wang et al. 2016; Lu et al. 2020). In fact, the information between different modalities are imbalanced and complementary. As shown in Fig. 1, we use texts to describe the image of the ice rink. Obviously, images often contain more details which have not been described by textual descriptions, such as woods and houses marked with red boxes. On the other hand, as shown at the bottom of Fig. 1, the textual descriptions contain more semantic information while image can not be demonstration, e.g., the texts marked with red color. Therefore, some fine-grained image details can not be exactly aligned to the textual descriptions and vice versa (Zhang et al. 2020). The existing methods simple building one common space would loses individual and useful modality-specific characteristics, which cannot fully exploit the intrinsic information within each modality.

To address this problem, in this paper, we propose a modality-specific matrix factorization hashing, which learns modality-specific semantic representation and then align them in the semantic space. As shown in Fig. 2, it includes off-line training and online testing stages. In the training, we learn individual semantic representations V_1 and V_2 for different modalities, respectively. Then, we use the orthogonal rotation transform to align the modality-specific



0.58	0.41	0.03	-0.51
-0.67	-0.67	0.67	0.60
-0.19	-0.30	-0.34	-0.22
0.22	0.34	-0.51	0.64
0.99	0.51	0.38	0.60

-0.11	-0.13	0.05
-0.04	0.05	-0.02
0.01	0.01	0.11
-0.03	-0.11	0.06
-0.06	0.08	0.10

-0.02	0.77	0.37	-0.06
0.26	0.96	0.03	0.17
-0.45	0.47	0.47	0.98

-0.47	-0.14	0.25
-0.24	-0.61	0.15
0.24	-0.19	0.51

0.06	-0.39	0.08	-0.17
0.09	-0.07	0.01	0.07
-0.13	-0.12	0.08	-0.23

0.05	0.07	0.37
0.08	-0.82	0.13
0.37	-0.11	0.25

0.46	-0.12	0.18
-0.22	0.53	0.07
-0.17	0.04	0.06

0.54	0.35	0.32	-0.46
-0.32	0.43	0.36	0.38
0.52	-0.51	0.37	0.26
0.46	0.25	-0.46	0.51

0.07	0.06	0.01
-0.06	-0.07	-0.01
0.05	0.12	0.06
0.14	-0.13	0.07

cat	duck	dog	cat
0	1	0	0
0	0	1	0
1	0	0	1

-0.01	-0.01	-0.01
-0.01	0.05	-0.04
0.01	0.01	0.05

1	-1	-1	1
-1	-1	-1	-1
-1	-1	1	-1

-0.56	-0.45	0.50	-0.36	0.03

-0.48	0.22	0.53	-0.58

1
-1
-1

1
1
-1

Fig. 2 The flowchart of proposed method. It includes offline training and online testing. For the offline stage, modality-specific matrix factorization on different modalities and alignment are incorporated to learn hash codes. For the online stage, the hash code of the query can be generated by the learned hash function, then cross-modal retrieval can be done by computing the similarity using XOR operation

representation V_1 and V_2 according to the correlation between different modalities. Finally, we embed the class label into the hash codes learning through the semantic latent space, and obtain directly the hash codes with closed solutions. The experimental results show that the learned hash codes are more discriminative compared to most existing methods. The contributions of the proposed method are summarized as follows:

- We propose modality-specific matrix factorization hashing via alignment, which learns the modality-specific representation for each modality and aligns them in the latent semantic space.
- We embed the hashing code learning through the semantic latent space, and obtain a closed solution for hash code learning.
- Extensive experiments on three datasets demonstrate the proposed method achieved better performance than most existing methods.

The rest of paper is organized as follows. We briefly review the related works in Sect.2 and elaborate the proposed method in Sect.3. Then, we present the optimization algorithm in Sect.4. After that, we conduct experiments in Sect.5 and give analysis in Sect.6. Finally, we conclude this work in Sect. 7.

2 Related works

Ding et al. (2014) first introduce the collective matrix factorization into cross-modal retrieval and proposed collective matrix factorization hashing (CMFH), which uses collective matrix factorization to learn unified hash codes for different modalities. Zhou et al. (2014) proposed a latent semantic sparse hash algorithm (LSSH), which learns the latent representations spaces by combining sparse coding and matrix factorization, and then merges the learned latent features to generate unified hash codes. Tang et al. (2016) extends CMFH and introduces manifold learning to learn more effective hash codes by preserving both inter-modal and intra-modal similarity. Yao Tao et al. (2019a) proposed a efficient discrete supervised hash algorithm (EDSH), which seamlessly integrates collective matrix factorization on heterogenous features and semantic embedding with class labels into hash codes. Although the correlation can be exploited by collective matrix factorization, however, they ignore individual information within each modality.

Recently, Peng et al. (2018) proposed a modality-specific cross-modal similarity measurement (MCSM), which constructs independent semantic space for each modality and directly computes the cross-modal similarity under the end-to-end framework. Yao Tao et al. (2019b) proposed a discrete semantic alignment hash algorithm (DSAH), which employed the attribute of image modality to align the semantic information with text modality, so as to exploit the intrinsic correlations among multiple modalities.

Inspired by those works, this paper focuses on learning modality-specific latent semantic representation and then align them according to the correlation between different modalities. Different from works (Peng et al. 2018; Yao Tao et al. 2019b), we achieve this idea under the principle of alignment.

3 Proposed method

3.1 Notations

In this work, we focus on two modalities, i.e., image modality and text modality. We denote the N image-text pairs as $X = \{X^{(1)}, X^{(2)}\} = \{\mathrm{x}_i^{(1)}, \mathrm{x}_i^{(2)}\}_{i=1}^{N}$, denote the associated semantic class label as $L = \{\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_N\} \in \{0, 1\}^{c \times N}$, where c is the total class number, and $L_{ij} = 1$ if x_i belongs to the j-th semantic category and 0 otherwise. $B = \{-1, 1\}^{k \times N}$ denotes the hash codes for the N image-text pairs, and k is the length of hash codes. Furthermore, we define the linear hash functions as follows,

$$\mathbf{h}^{(1)}(\mathrm{x}_i^{(1)}) = sgn(W_1 \mathrm{x}_i^{(1)}), \qquad \mathbf{h}^{(2)}(\mathrm{x}_i^{(2)}) = sgn(W_2 \mathrm{x}_i^{(2)})$$

where W_1 and W_2 denote the hash functions that map the data points into the Hamming space, $sgn(\cdot)$ is an element-wise sign function. Following existing works (Zhang and Li 2014; Ma et al. 2016), we assume that the feature vectors are zero-centered, i.e., $\sum_{i=1}^{N} \mathrm{x}_{ij}^{(1)} = 0$, $\sum_{i=1}^{N} \mathrm{x}_{ij}^{(2)} = 0$.

3.2 Formulation

3.2.1 Learning modality-specific representation

The collective matrix factorization hashing decomposes the different modality representation matrix into a common representation matrix, which inherently sacrifices the representation accuracy and loses modality-specific information for each modality. To solve this problem, we learn the modality-specific representations individually for each modality and then align them in the semantic space. Thus, learning the modality-specific representations can be formulated below for each modality,

$$\mathop{arg\,min}_{U_i, V_i} \alpha_i \|X^{(i)} - U_i V_i\|_F^2 + \mu\left(\|U_i\|_F^2 + \|V_i\|_F^2\right) \tag{1}$$

where $i = 1, 2$, $U_1 \in \mathbb{R}^{d_1 \times k}$ and $U_2 \in \mathbb{R}^{d_2 \times k}$, $V_1 \in \mathbb{R}^{k \times N}$ and $V_2 \in \mathbb{R}^{k \times N}$ are the modality-specific representations, α_i and μ are regularization parameters.

3.2.2 Aligning modality-specific representation

To model the correlations between different modalities, we align the modality-specific representations from different modalities by solving following problem,

$$\arg\min_{R} \beta \| V_1 - RV_2 \|_F^2 \quad s.t. \quad RR^T = I \tag{2}$$

where $R \in \mathbb{R}^{k \times k}$, and I denotes the identity matrix, β is a regularization parameter.

3.2.3 Learning hash functions

To deal with out-of-sample, we learn modality-specific hash functions for different modalities by following formulation,

$$\arg\min_{W_i} \gamma \| V_i - W_i X^{(i)} \|_F^2 + \mu \| W_i \|_F^2 \tag{3}$$

where $i = 1, 2$, $W_1 \in \mathbb{R}^{k \times d_1}$ and $W_2 \in \mathbb{R}^{k \times d_2}$ are the hash functions for the image and text modalities, γ is a regularization parameter.

3.2.4 Learning hash codes

Many existing supervised cross-modal hashing methods improve discrimination of the hash codes through constructing a $N \times N$ similarity matrix , which leads large computational cost with increase of N. On the other hand, the transformation from class label into similarity matrix also results in information loss. To remedy this problem, we directly embed the class label into the hash codes learning via latent semantic space as following,

$$\arg\min_{P} \eta \| L - PB \|_F^2 + \mu \| P \|_F^2 \tag{4}$$

where latent semantic space $P \in \mathbb{R}^{c \times k}$ bridges the semantic correlations between class labels and hash codes. $B_{ij} = 1$ indicates that the *j-th* data point contains the semantics of *i-th* data point, otherwise $B_{ij} = -1$.

3.2.5 Bridging the gap

To narrow the semantic gap between hash codes and modality-specific representations, we further define the following formulation,

$$\begin{aligned} &\arg\min R_i \| B - R_i V_i \|_F^2 \\ s.t. \quad & R_i R_i^T = I, B \in \{-1, 1\}^{k \times N} \end{aligned} \tag{5}$$

where $R_i \in \mathbb{R}^{k \times k}$, $i = 1, 2$. Through this formulation, the hash codes can be obtained directly during training stage and the quantization errors can also be reduced.

3.2.6 Objective function

Combining the terms given from Eq. (1) to Eq. (5), we obtain the whole objective function as below,

$$\begin{aligned} \arg\min & \sum_{i=1}^{2} \{ \| B - R_i V_i \|_F^2 + \alpha_i \| X^{(i)} - U_i V_i \|_F^2 \\ & + \gamma \| V_i - W_i X^{(i)} \|_F^2 \} + \beta \| V_1 - RV_2 \|_F^2 \\ & + \eta \| L - PB \|_F^2 + \mu Q \\ s.t. \quad & B \in \{-1, 1\}^{k \times N}, \quad RR^T = I, \quad R_i {R_i}^T = I \end{aligned} \tag{6}$$

where $Q = \sum_{i=1}^{2} \{ \| U_i \|_F^2 + \| V_i \|_F^2 + \| W_i \|_F^2 \} + \| P \|_F^2$.

4 Algorithm

4.1 Optimization

The problem (6) is no-convex with all the variables. Fortunately, it is convex with respect to one variable when the other variables are fixed. Therefore, we utilize an alternative optimal algorithm to solve the subproblems with respect to each variable.

U_i-**Step:** Fixed the other variables and dropping the irrelevant terms with respect to $U_i, i = 1, 2$, we obtain the following subproblem,

$$\arg\min_{U_i} \alpha_i \| X^{(i)} - U_i V_i \|_F^2 + \mu (\| U_i \|_F^2)$$

Obviously, this is a convex quadratic optimization problem, we can obtain the closed solution for U_i as following,

$$U_i = X^{(i)} V_i^T \left(V_i V_i^T + \frac{\mu}{\alpha_i} I \right)^{-1}, i = 1, 2. \tag{7}$$

P-**Step:** Fixed the other variables and setting the derivation of Eq. (6) with respect to P as zero, we have,

$$P = LB^T \left(BB^T + \frac{\mu}{\eta} I \right)^{-1} \tag{8}$$

V_i-**Step:** Fixed the other variables, we derive following subproblem with respect to V_1,

$$\begin{aligned} \arg\min_{V_1} & \alpha_1 \| X^{(1)} - U_1 V_1 \|_F^2 + \| B - R_1 V_1 \|_F^2 \\ & + \gamma \| V_1 - W_1 X^{(1)} \|_F^2 + \beta \| V_1 - RV_2 \|_F^2 \end{aligned} \tag{9}$$

Setting the derivation of Eq. (9) with respect to V_1 as zero, we have,

$$\begin{aligned} & \alpha_1 U_1^T U_1 V_1 + ((\beta + \gamma + \mu) I + R_1^T R_1) V_1 \\ & \quad - (\alpha_1 U_1^T + \gamma W_1) X^{(1)} - \beta RV_2 - R_1^T B = 0 \end{aligned}$$

Thus, we can obtain a closed solution for V_1 as following,

$$V_1 = \left(\alpha_1 U_1^T U_1 + (\beta + \gamma + \mu)I + R_1^T R_1\right)^{-1} \left(\alpha_1 U_1^T X^{(1)} + \beta R V_2 + \gamma W_1 X^{(1)} + R_1^T B\right) \tag{10}$$

Similar to V_1 , the solution for V_2 is as below,

$$V_2 = \left(\alpha_2 U_2^T U_2 + \beta R^T R + (\gamma + \mu)I + R_2^T R_2\right)^{-1} \left(\alpha_2 U_2^T X^{(2)} + \beta R^T V_1 + \gamma W_2 X^{(2)} + R_2^T B\right) \tag{11}$$

R-**Step:** Fixed the other variables, we derive the subproblem respect to R,

$$arg \min_{R} \|V_1 - RV_2\|_F^2, \qquad s.t. RR^T = I \tag{12}$$

Problem (12) is a classical orthogonal procrustes problem (Schönemann 1966), which can be solved by singular value decomposition (SVD). Specifically, the SVD of $V_1 V_2^T$ is firstly computed as $V_1 V_2^T = S\Omega\tilde{S}^T$, where $S, \tilde{S} \in \mathbb{R}^{k\times k}$ are orthogonal matrix, and $\Omega \in \mathbb{R}^{k\times k}$ is diagonal matrix, the columns of S and $\tilde{S}$ are singular vectors and the diagonal elements of Ω are singular values. And then the orthogonal matrix R can be updated by $R = \tilde{S}S^T$. The update of R_1, R_2 are similar.

B-**Step:** Fixed the other variables and dropping the irrelevant terms to B, we obtain

$$\begin{aligned} arg \min_{B} & \|B - R_1 V_1\|_F^2 + \|B - R_2 V_2\|_F^2 \\ & + \eta \|L - PB\|_F^2 \\ s.t. \quad & B \in \{-1, 1\}^{k\times N} \end{aligned} \tag{13}$$

Equation (13) is equivalent to

$$\begin{aligned} \min_{B} & tr((B^T B) - 2(V_1^T R_1^T B) + \eta(B^T P^T PB) \\ & - 2\eta(L^T PB) - 2(V_2^T R_2^T B)) \end{aligned}$$

Since $tr(B^T B)$ and $tr(B^T P^T PB)$ are constants, we can obtain a closed solution for B as following

$$B = sgn\left(R_1 V_1 + R_2 V_2 + \eta P^T L\right) \tag{14}$$

W_i-**Step:** Fixed the other variables and dropping the irrelevant terms to $W_i, i = 1, 2$, we obtain

$$arg \min_{W_i} \gamma \|V_i - W_i X^{(i)}\|_F^2 + \mu \|W_i\|_F^2 \tag{15}$$

Setting the derivation of Eq. (15) with respect to W_i equal zero, we have

$$W_i = V_i X^{(i)T} \left(X^{(i)} X^{(i)T} + \frac{\mu}{\gamma} I\right)^{-1} \tag{16}$$

Repeating the above steps until the stop condition is reached, the optimal solution of the variables can be obtained. It is worth noting that each valuable has a closed solution, from which discrete hash codes can be obtained directly. The whole procedure is summarized in algorithm 1.

```
Algorithm 1 Modality-specific Matrix Factorization Hashing (MsMFH)
Input: Training data $\{X^{(1)}, X^{(2)}\}$ and the corresponding class label matrix $L$, the hash codes length $k$, and parameters $\alpha, \beta, \gamma, \eta, \mu$.
Initializing $B$, $V_1$, $V_2$, $R$, $R_1$, $R_2$, and hash functions $W_1$, $W_2$, randomly.
for $i = 1$ to $Iter$ do
    Update $U_i, i = 1, 2$ by equation (7) with the other variables fixed,
        $U_i = X^{(i)} V_i^T \left(V_i V_i^T + \frac{\mu}{\alpha_i} I\right)^{-1}$
    Update $P$ by equation (8) with other valuables fixed,
        $P = LB^T \left(BB^T + \frac{\mu}{\eta} I\right)^{-1}$
    Update $V_1$ by equation (10) with other variables fixed,
        $V_1 = \left(\alpha_1 U_1^T U_1 + (\beta+\gamma+\mu)I + R_1^T R_1\right)^{-1} \left(\alpha_1 U_1^T X^{(1)} + \beta R V_2 + \gamma W_1 X^{(1)} + R_1^T B\right)$
    Update $V_2$ by equation (11) with other variables fixed,
        $V_2 = \left(\alpha_2 U_2^T U_2 + \beta R^T R + (\gamma+\mu)I + R_2^T R_2\right)^{-1} \left(\alpha_2 U_2^T X^{(2)} + \beta R^T V_1 + \gamma W_2 X^{(2)} + R_2^T B\right)$
    Update $R$, $R_1$, $R_2$ using $R = \tilde{S}S^T$, $R_1 = \tilde{S}_1 S_1^T$, $R_2 = \tilde{S}_2 S_2^T$ with other variables fixed, respectively.
    Update the hash codes $B$ by equation (14) with other variables fixed,
        $B = sgn\left(R_1 V_1 + R_2 V_2 + \eta P^T L\right)$
    Update the hash functions $W_i, i = 1, 2$ by equation (16) with other variables fixed,
        $W_i = V_i X^{(i)T} \left(X^{(i)} X^{(i)T} + \frac{\mu}{\gamma} I\right)^{-1}$
end for
Output: The hash functions $W_1$, $W_2$, the matrix $R_1$, $R_2$.
```

4.2 Out-of-sample extension

Based on the output of algorithm 1, we can easily deal with the out-of-sample extension. For example, given a query $x_q^{(1)}$ from image modality, we can obtain the hash codes as follows,

$$B_q = sgn(R_1 W_1 x_q^{(1)})$$

Similarly, we can compute the hash codes given the query $x_q^{(2)}$ from text modalities below,

$$B_q = sgn(R_2 W_2 x_q^{(2)})$$

4.3 Complexity analysis

The computational load mainly includes the following parts in the training: Eqs. (7) and (8) are $O(d_1kN + Nk^2 + k^3 + d_1k^2)$, $O(d_2kN + Nk^2 + k^3 + d_2k^2)$, and $O(ckN + Nk^2 + k^3 + ck^2)$, respectively; Eqs. (10) and (11) are $O(k^2d_1 + 2k_3 + 2kNd_1 + 2k^2N)$ and $O(k^2d_2 + 3k_3 + 2kNd_2 + 2k^2N)$, respectively; the update of R, R_1 and R_2 are $O(k^3 + k^2N)$, $O(k^3 + k^2N)$, and $O(k^3 + k^2N)$, respectively; Eq. (14) is $O(2k^2N + kcN + kN)$; Eq. (16) are $O(kd^1N + d_1^2N + kd_1^2 + d_1^3)$ and $O(kd^2N + d_2^2N + kd_2^2 + d_2^3)$, respectively. Usually, the training number N is much greater than k, d_1, d_2 and c, thus the complexity in each iteration is linear to the training size N. Given the number of iterations $Iter$, the overall training computational complexity of MsMFH is $O(N)$.

5 Experiments and results

To evaluate the effectiveness of the proposed method, we conduct experiments on three datasets. First, we introduce the datasets and evaluation in Sect.5.1, and then present the implementation details and compared methods in Sect.5.2. Finally, we show the experimental results in Sect.5.3.

5.1 Datasets and evaluation

5.1.1 Datasets

We conduct experiments on the Wikipedia (Rasiwasia et al. 2010), Mirflickr25k (Huiskes and Lew 2008) and NUS-WIDE (Chua et al. 2009) datasets, which are shown in Table 1. The Wikipedia dataset consists of 10 categories, 2866 instances (image-text pairs), in which 2173 image-text pairs are randomly selected for training and the rest of 693 image-text pairs are selected for testing. The Mirflickr25k includes 20,015 image-text pairs, in which the 4015 image-text pairs are randomly selected for training and the 16,000 image-text pairs are selected for testing. The NUS-WIDE includes 10 categories and contains 186,776 image-text pairs, in which the 184,710 image-text pairs are randomly selected for training and the 1867 image-text pairs are selected for testing.

Table 1 The details of three datasets, where "/" in column "Instances" represents the number of training/test image-text pairs

Dataset	Instances	Categories	Image feature	Text feature
Wikipedia	2173/693	10	4096d VGG	10d BoW
Mirflickr25k	4015/16000	24	4096d VGG	1386d BoW
NUS-WIDE	184710/1867	10	500d VGG	1000d BoW

As shown in Table 1, the 4096-dimension VGG are selected as the image features for the Wikipedia and Mirflickr25k dataset, while 500-dimension VGG features are selected as image features for NUS-WIDE dataset. The text features are 10-dimensional topics vector in Wikipedia, and 1386-dimensional topics vector in Mirflickr25k, and 1000-dimensional topics vector in NUS-WIDE dataset. For all the compared methods, the dataset partition and feature extraction methods are the same.

5.1.2 Evaluation

The cross-modal retrieval performance is evaluated by the mean average precision (mAP), which is the mean of average precision. The average precision (AP) for each query is defined as follows,

$$\mathrm{AP} = \frac{1}{l} \sum_{i=1}^{m} p(i)\delta(i)$$

where m is the number of returned results, l is the total number of semantically related the query, $p(i)$ denotes the precision of the top i returned results, and $\delta(i)$ is an indicator function. $\delta(i) = 1$ if the i-th entity is relevant to the query, otherwise $\delta(i) = 0$. Generally, the mAP is larger, the retrieval performance is better. We conduct two tasks includes retrieving text using image as query (Img2Txt) and retrieving image using text as query (Txt2Img), and report the mAP. Besides, we report the precision-recall performance on the three datasets.

Table 2 The mAP@100 scores comparison on Wikipedia dataset

Methods	Img2Txt				Txt2Img			
	8	16	32	64	8	16	32	64
CCA	0.214	0.190	0.165	0.159	0.423	0.319	0.230	0.189
IMH	0.260	0.217	0.195	0.172	0.366	0.319	0.287	0.252
STMH	0.241	0.254	0.279	0.292	0.272	0.326	0.364	0.380
CMFH	0.165	0.169	0.174	0.180	0.215	0.225	0.227	0.228
SMFH	0.159	0.159	0.160	0.163	0.241	0.240	0.255	0.270
SCRATCH-o	0.192	0.214	0.244	0.270	0.306	0.374	0.480	0.538
SCM-orth	0.154	0.167	0.181	0.184	0.339	0.353	0.357	0.359
EDSH	0.373	0.387	0.394	0.395	0.558	0.566	0.566	0.573
MsMFH(ours)	0.383	0.395	0.404	0.409	0.589	0.608	0.618	0.622

5.2 Implementation details and compared methods

For fair comparison, we implement the source codes of the EDSH and initialize the parameters according to the paper suggestion. For our proposed method, the mini-batch is size to 128 and the parameters are empirically set as $\alpha_i = 1$, $\gamma = 10$, $\beta = 2$, $\eta = 10$, and $\mu = 5$. More analysis with the parameter setting are presented in Sect.6. All the experiments are run five times, and the average values are reported as the final results. We selected following representative methods for comparison.

- CCA (Rasiwasia et al. 2010) : It learns linear projection by maximizing the correlation between pairwise data from different modalities.
- IMH (Song et al. 2013) : It explores the correlations among different modalities and learns a common hamming space.
- STMH (Wang et al. 2015) : It learns discrete hash codes by considering latent semantic information in coding procedure.
- CMFH (Ding et al. 2014) : It uses collective matrix factorization to learn unified hash codes for different modalities.
- SMFH (Tang et al. 2016) : It preserves the similarities among multi-modal original features through a graph regularization.
- SCRATCH-o (Chen et al. 2019) : It utilizes collective matrix factorization on original features and learn the latent representations in a shared latent space with label semantic embedding.
- SCM-orth (Zhang and Li 2014) : It seamlessly integrates semantic labels into the hashing learning procedure for large-scale data modeling.
- EDSH (Yao Tao et al. 2019a) : It seamlessly integrates the collective matrix factorization for heterogeneous features and semantic embedding with class labels to learn hash codes.

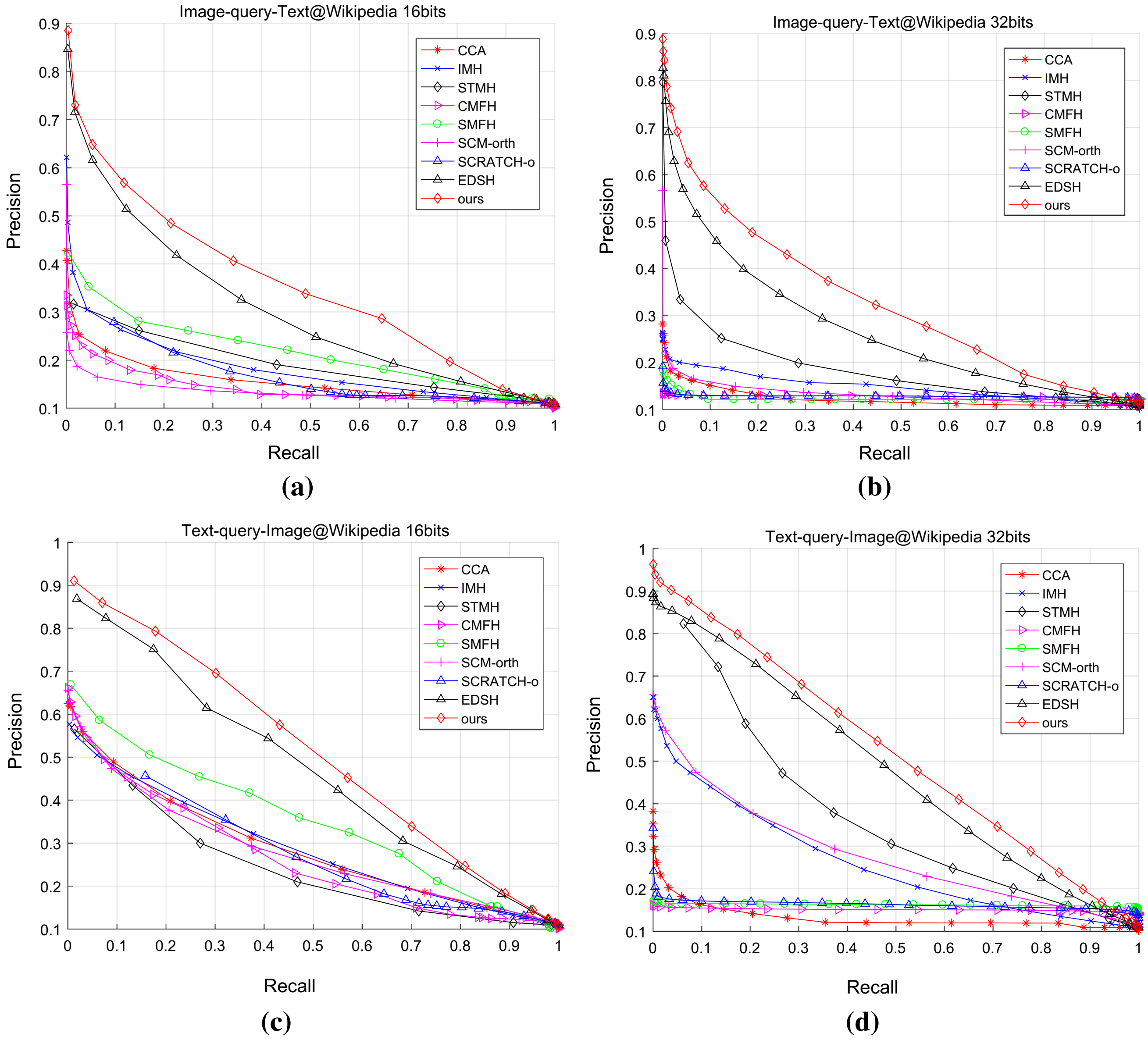
5.3 Experimental results

5.3.1 Mean average precision

Tables 2, 3 and 4 report the mAP scores of different methods with respect to different hash codes lengths on the three dataset. From that, we have following observations. On the wikipedia dataset, the performances of all the methods are worse than that of the other two datasets. It might be the training data is small than other two datasets. For wikipedia and NUS-WIDE dataset, the performance of Txt2Img is better than that of the Img2Txt task. Intuitively, the reason behind this is that the text modality is suitable for semantic representation. Compared to other methods, the performance of MsMFH continuously increases, with the increase of the hash code length. On the three dataset, our MsMFH approach yields much better mAP results than that of the compared methods in the most experiments, which demonstrates its effectiveness. Compared to EDSH, MsMFH outperforms up to 3% on the mAP when the hash code lengths varying from 8 bits to 64 bits. This demonstrates the efficiency of MsMFH.

5.3.2 Precision-recall curves

Figures 3, 4 and 5 show the precision-recall curves with the hash code length varying from 16 and 32 bits on the three datasets. From that, we can conclude that MsMFH achieves best performance compared to all baseline methods consistently when the recall is relatively small.



(a) **(b)** **(c)** **(d)**

Fig. 3 Precision-Recall curves on Wikipedia dataset when hash codes are 16 and 32 bits

Table 3 The mAP@100 scores comparison on Mirflickr25k dataset

Methods	Img2Txt				Txt2Img			
	8	16	32	64	8	16	32	64
CCA	0.596	0.595	0.596	0.596	0.583	0.583	0.606	0.601
IMH	0.626	0.612	0.597	0.585	0.662	0.646	0.625	0.608
STMH	0.617	0.633	0.637	0.655	0.590	0.626	0.639	0.637
CMFH	0.560	0.560	0.562	0.562	0.563	0.566	0.566	0.568
SMFH	0.564	0.566	0.565	0.567	0.578	0.577	0.582	0.583
SCRATCH-o	0.567	0.576	0.579	0.581	0.575	0.592	0.604	0.614
SCM-orth	0.583	0.573	0.567	0.564	0.675	0.646	0.675	0.591
EDSH	0.845	0.865	0.885	0.888	0.795	0.801	0.802	0.809
MsMFH(ours)	0.854	0.883	0.895	0.899	0.829	0.836	0.840	0.847

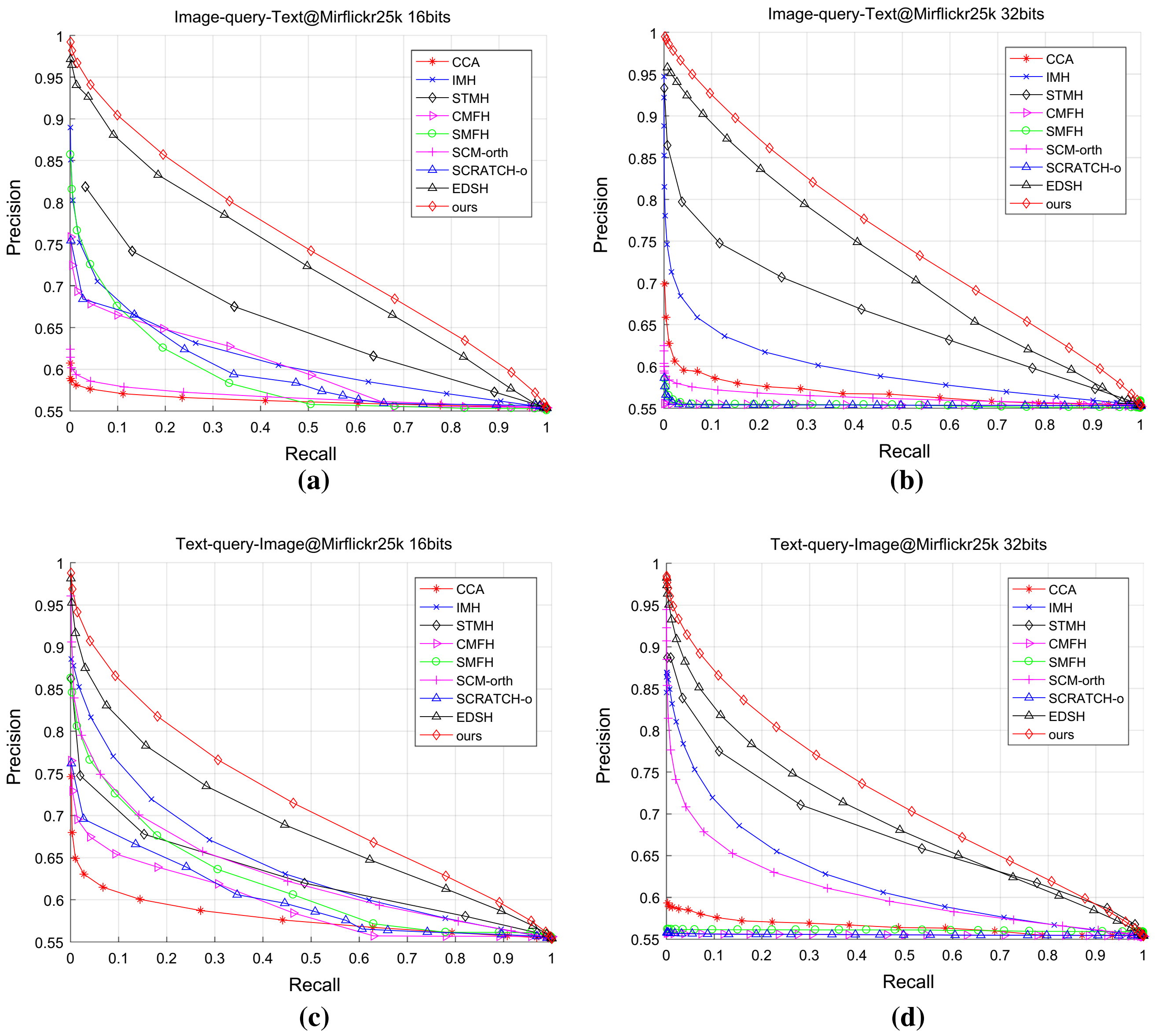
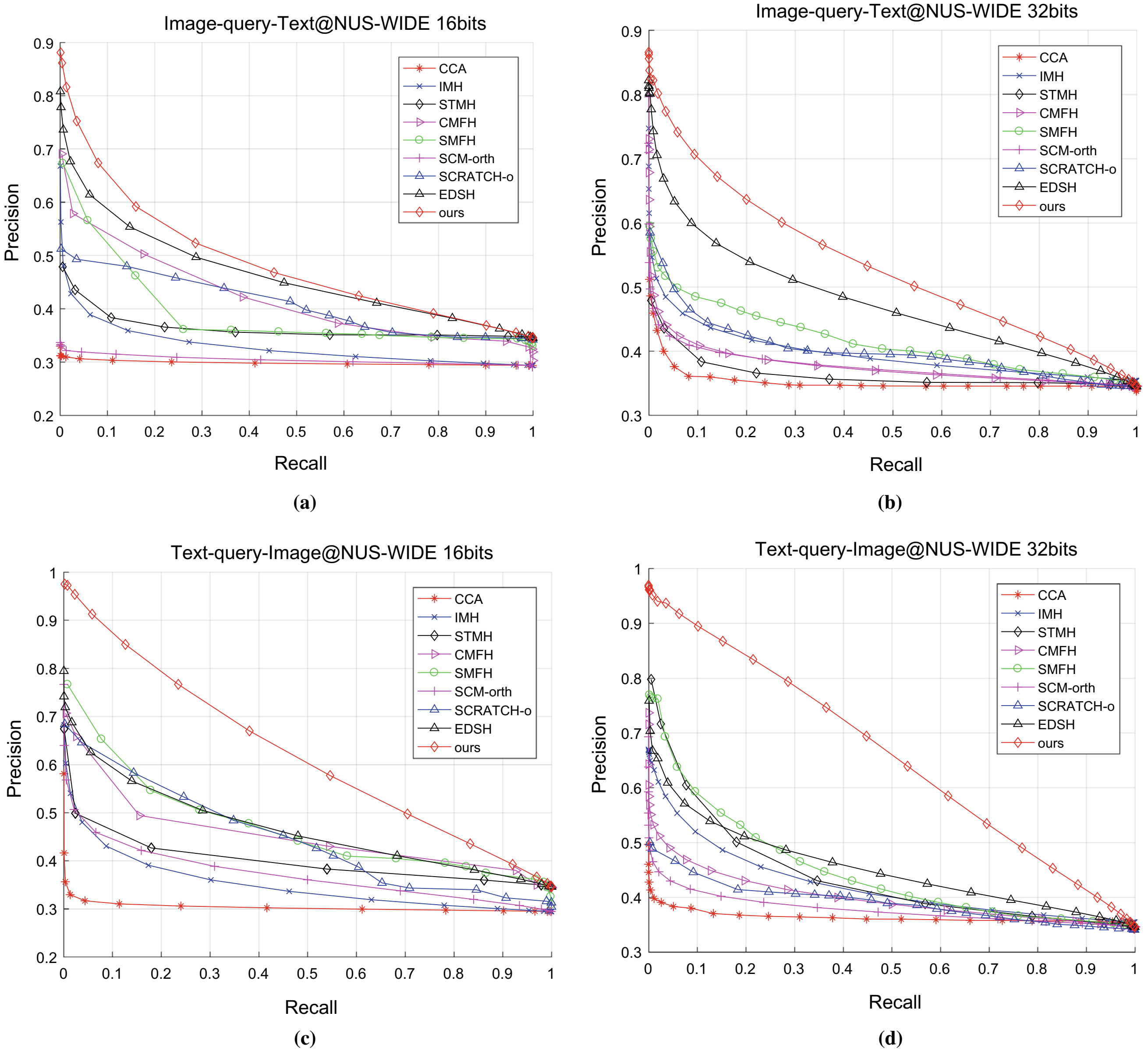


Fig. 4 Precision-Recall curves on Mirflickr25k dataset when hash codes are 16 and 32 bits

Table 4 The mAP@100 scores comparison on NUS-WIDE dataset

Methods	Img2Txt				Txt2Img			
	8	16	32	64	8	16	32	64
CCA	0.359	0.353	0.351	0.349	0.344	0.345	0.344	0.342
IMH	0.481	0.482	0.484	0.487	0.563	0.544	0.594	0.612
STMH	0.347	0.346	0.346	0.390	0.390	0.405	0.408	0.432
CMFH	0.442	0.473	0.499	0.503	0.476	0.518	0.566	0.583
SMFH	0.432	0.448	0.457	0.463	0.415	0.413	0.411	0.402
SCRATCH-o	0.471	0.494	0.519	0.421	0.507	0.536	0.555	0.554
SCM-orth	0.428	0.405	0.388	0.375	0.422	0.395	0.378	0.367
EDSH	0.514	0.521	0.539	0.554	0.634	0.656	0.669	0.683
MsMFH(ours)	0.552	0.564	0.589	0.607	0.679	0.695	0.712	0.716



(a)

(b)

(c)

(d)

Fig. 5 Precision-Recall curves on NUS-WIDE dataset when hash codes are 16 and 32 bits

5.3.3 Top-10 retrieval results

Figure 6 shows the top-10 retrieval results by the CCA, EDSH and MsMFH on Wikipedia dataset. From that, we can observe that CCA has more failure cases, which are marked with red color boxes compared to EDSH, and MsMFH. Although the proposed method is also unsuccessful in one case in the task of Txt2Img, the retrieved results is intuitively semantic correlation with the query. The reason behind this might be that the alignment of latent semantic representation is wrong.

6 Analysis and discussion

In this section, we will further discuss the proposed method from the learned representation, parameter selection and convergence.

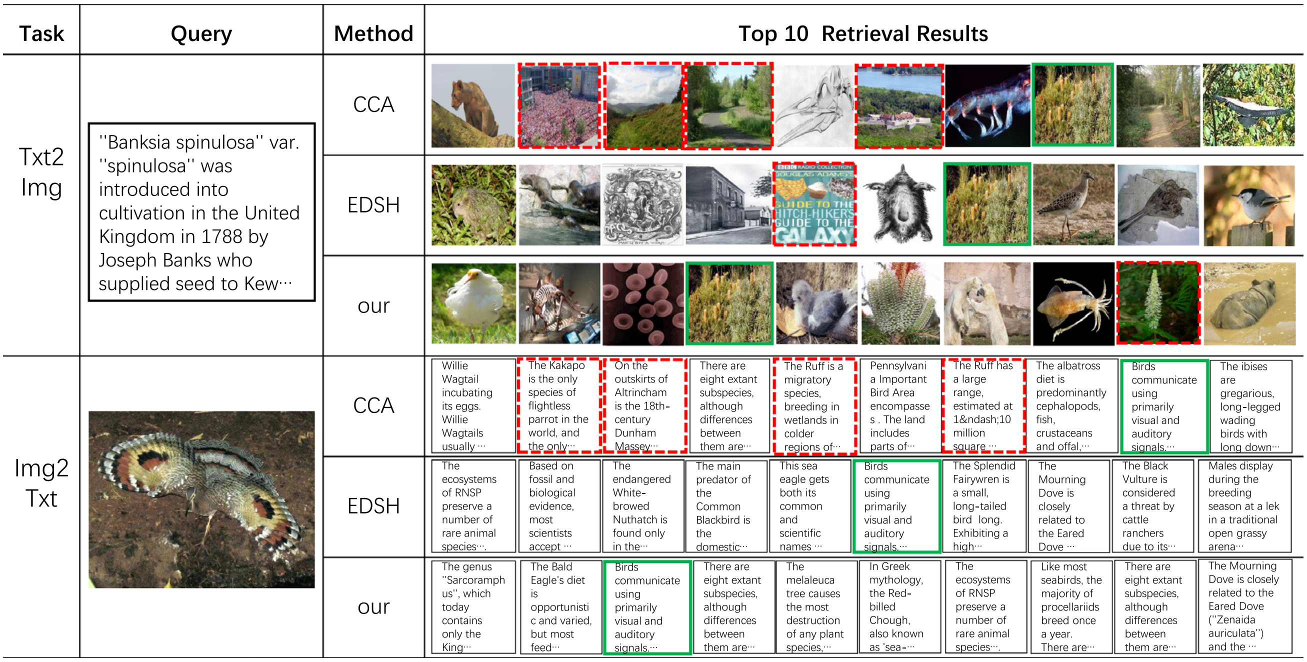

Task
Query
Method
Top 10 Retrieval Results
Txt2 Img
"Banksia spinulosa" var. "spinulosa" was introduced into cultivation in the United Kingdom in 1788 by Joseph Banks who supplied seed to Kew···
CCA
EDSH
our
Img2 Txt
CCA
Willie Wagtail incubating its eggs. Willie Wagtails usually ···
The Kakapo is the only species of flightless parrot in the world, and the only···
On the outskirts of Altrincham is the 18th-century Dunham Massey···
There are eight extant subspecies, although differences between them are···
The Ruff is a migratory species, breeding in wetlands in colder regions of···
Pennsylvani a Important Bird Area encompasse s . The land includes parts of···
The Ruff has a large range, estimated at 1&ndash;10 million square ···
The albatross diet is predominantly cephalopods, fish, crustaceans and offal,···
Birds communicate using primarily visual and auditory signals.···
The ibises are gregarious, long-legged wading birds with long down···
EDSH
The ecosystems of RNSP preserve a number of rare animal species···.
Based on fossil and biological evidence, most scientists accept ···
The endangered White-browed Nuthatch is found only in the···
The main predator of the Common Blackbird is the domestic···
This sea eagle gets both its common and scientific names ···
Birds communicate using primarily visual and auditory signals.···
The Splendid Fairywren is a small, long-tailed bird long. Exhibiting a high···
The Mourning Dove is closely related to the Eared Dove ···
The Black Vulture is considered a threat by cattle ranchers due to its···
Males display during the breeding season at a lek in a traditional open grassy arena···
our
The genus ''Sarcoramph us'', which today contains only the King···
The Bald Eagle's diet is opportunisti c and varied, but most feed···
Birds communicate using primarily visual and auditory signals.···
There are eight extant subspecies, although differences between them are···
The melaleuca tree causes the most destruction of any plant species,···
In Greek mythology, the Red-billed Chough, also known as 'sea-···
The ecosystems of RNSP preserve a number of rare animal species···.
Like most seabirds, the majority of procellariids breed once a year. There are···
There are eight extant subspecies, although differences between them are···
The Mourning Dove is closely related to the Eared Dove (''Zenaida auriculata'') and the ···

Fig. 6 Top-10 retrieval results by the CCA, EDSH and MsMFH on Wikipedia dataset, the results with green boxes are correct, while those with red dotted boxes are wrong

6.1 Representation analysis

6.1.1 Representation visualization

To shown the latent semantic representation of EDSH and MsMFH, we conduct experiment on wikipedia dataset. We select three semantic categories, i.e., biology, geography and warfare. We learn the low-dimensional representation using EDSH and MsMFH, respectively. Then, we utilize the t-SNE to perform dimensionality reduction and present their two dimensional representations in Fig. 7. It can be seen that the latent semantic representation learned by MsMFH are more closer for each class than that of EDSH.

6.1.2 Representation alignment

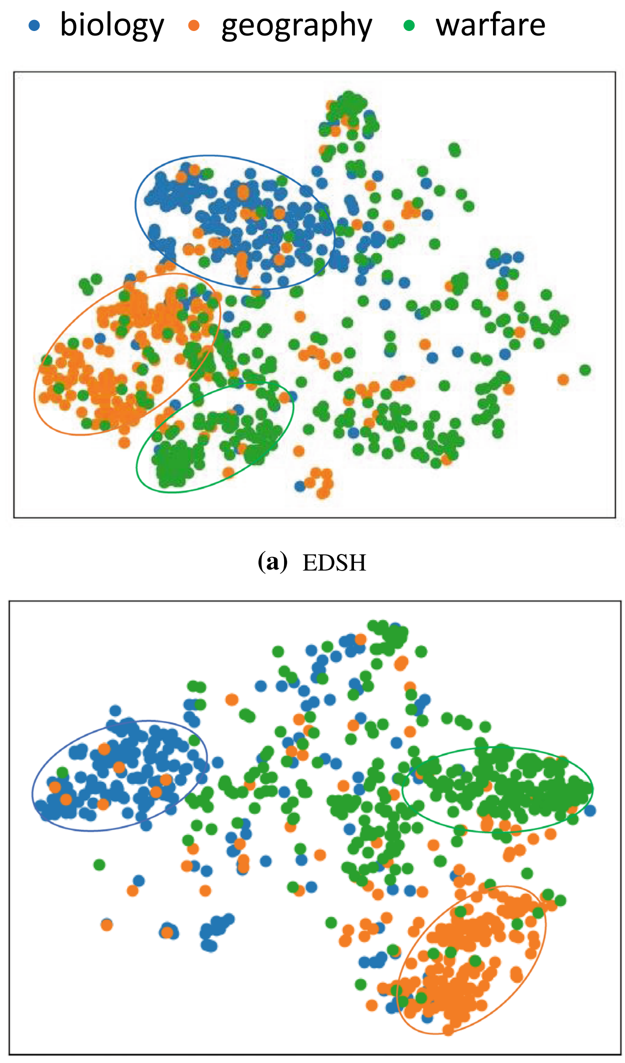
The alignment of latent semantic representation is very important in the proposed method. To demonstrate the effectiveness, we conduct experiments on wikipedia dataset with the same setting as part 6.1.1. As presented in Fig. 8, (a) is the image modality-specific representations, and (b) is the text modality-specific representations, (c) is the results after alignment using proposed method. From that, we can see that same semantic category are almost overlap after alignment.

6.2 Complexity analysis

In order to compare the training time, we run the methods five times in all the experiments, and report the average time as the final time, which are reported in Table 5. Because of IMH and SMFH spend a long time on NUS-WIDE datasets, we do not discuss their training time here. From Table 5, it can be seen that our MsMFH costs relatively less time than that of EDSH.

6.3 Convergence analysis

To show the convergence of the proposed method, we conduct experiments on the three datasets to compared with EDSH. The convergence curves on the three datasets are shown in Fig. 9. It can be observed that MsMFH method on three datasets converges slightly faster than that of EDSH.



(a) EDSH

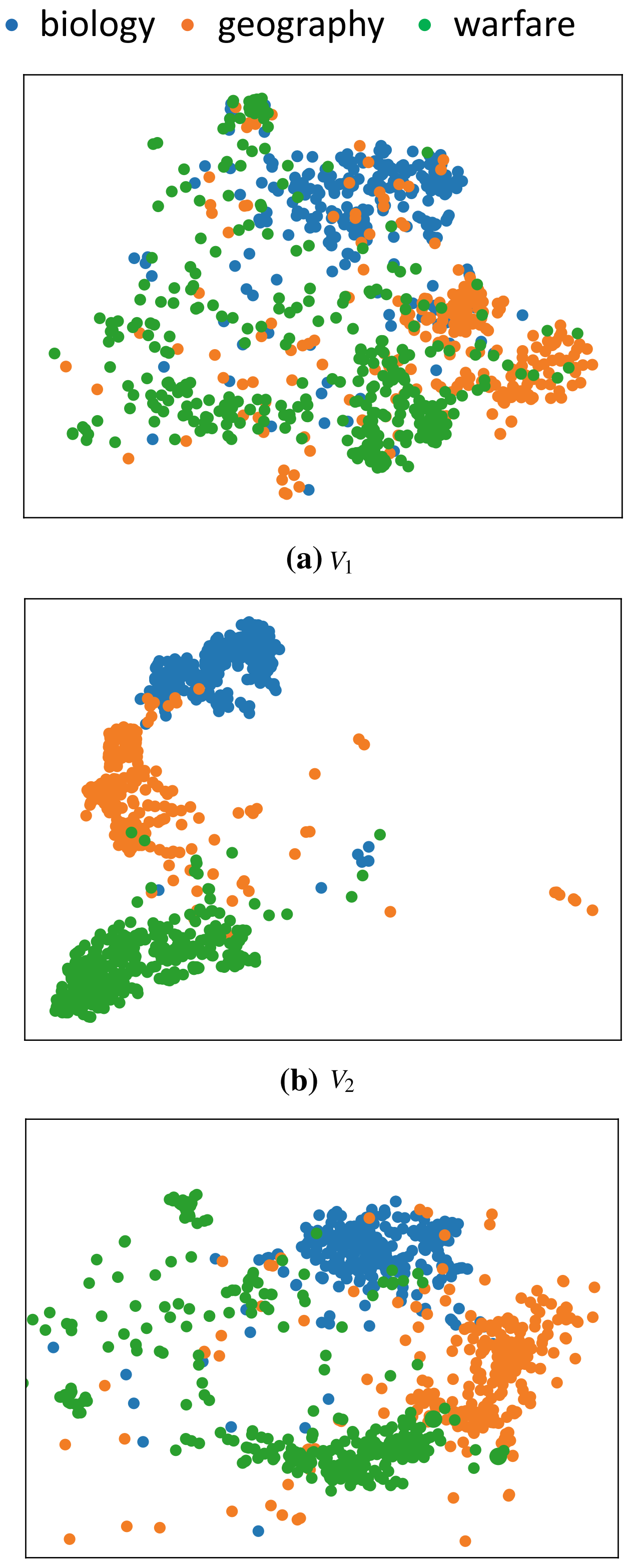
(b) MsMFH

Fig. 7 The t-SNE visualization of latent semantic representations for three semantic categories in the Wikipedia dataset with different colors representing different semantic categories

6.4 Parameter analysis

In this part, we analyze the parameter sensitivity of MsMFH via experiments, which are carried out by varying the value of one parameter with the others fixed. The parameters α_1 and α_2 balances the importance of different modalities. In fact, the two modalities are considered equally. Therefore, we empirically set $\alpha_1 = \alpha_2 = 1$.

Therefore, we only analyze the parameters of β, η, γ, μ. The mAP curves in the case of 32 bits with varying the values β, η, γ, μ are reported in Fig. 10. From those, we can observe:



(a) V_1

(b) V_2

(c) RV_2

Fig. 8 The alignment of latent semantic representations learned by the proposed method for three semantic categories in Wikipedia dataset

Table 5 Training time (seconds) comparison on the three datasets for the 32 bits

Methods	Wikipedia	Mirflickr25k	NUS-WIDE
CCA	110	2054	812
IMH	65	258	-
STMH	226	408	2190
CMFH	237	435	1292
SMFH	244	848	-
SCRATCH-o	881	1149	3210
SCM-orth	1535	1612	2042
EDSH	60	216	357
MsMFH (ours)	58	203	342

- For β and η, MsMFH obtains stable performance in wide range of from 0.001 to 100 on the three datasets.
- For the parameter γ, the mAP values vary dramatically in the range from 0.001 to 100 on the wikipedia dataset, this might be the the dataset size is smaller than other two datasets.
- For γ and μ, the performances of MsMFH are stable in wide range from 0.001 to 100 on the Mirflick25k and NUS-WIDE datasets.

7 Conclusion

In this paper, we propose a modality-specific matrix factorization hashing, i.e., Modality-specific Matrix Factorization Hashing for Cross-modal Retrieval (MsMFH). It first learns modality-specific representation, then aligns them in the semantic space, following by semantic embedding with class labels to improve the discrimination of hash codes. We conduct experiments on three public real-world datasets. From the results, we find that the proposed method achieved better retrieval performances compared to existing methods, the learned representation are more discriminative. At the same time, the analysis of complexity and convergence for the proposed method also show its advantages compared with existing methods. In the future, we will discuss more efficient alignment method.

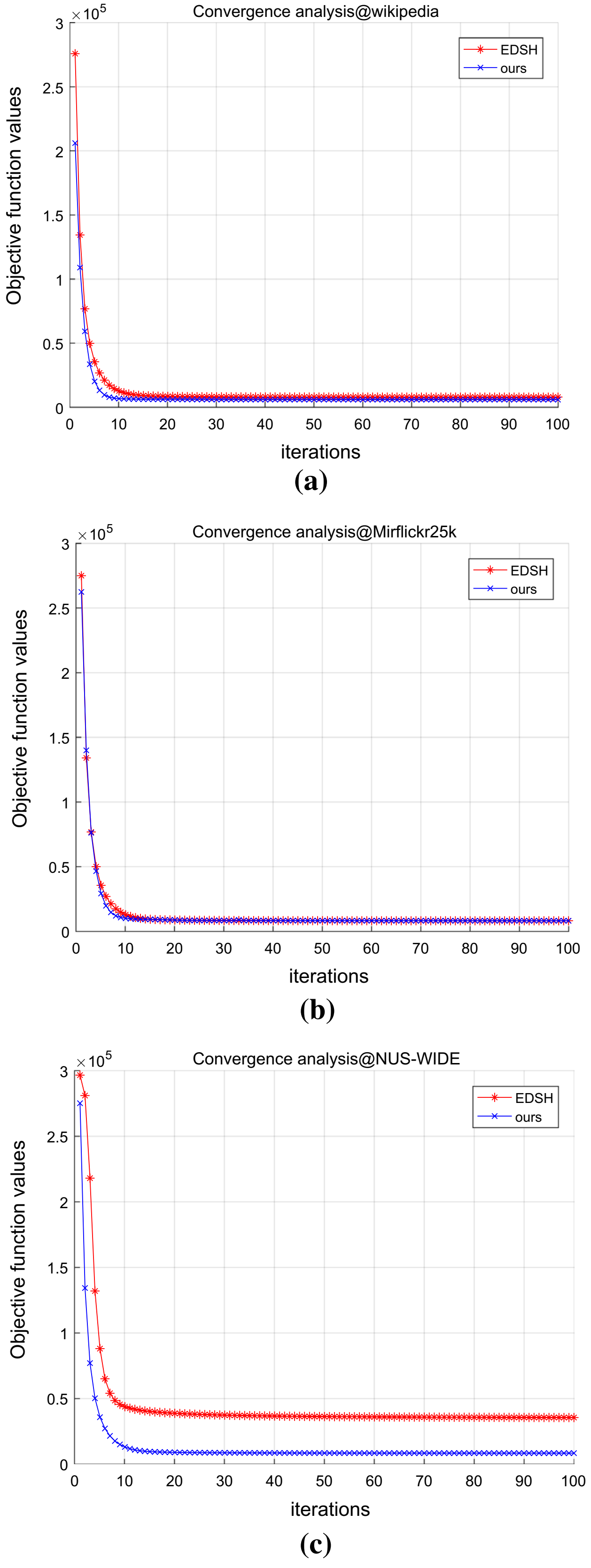
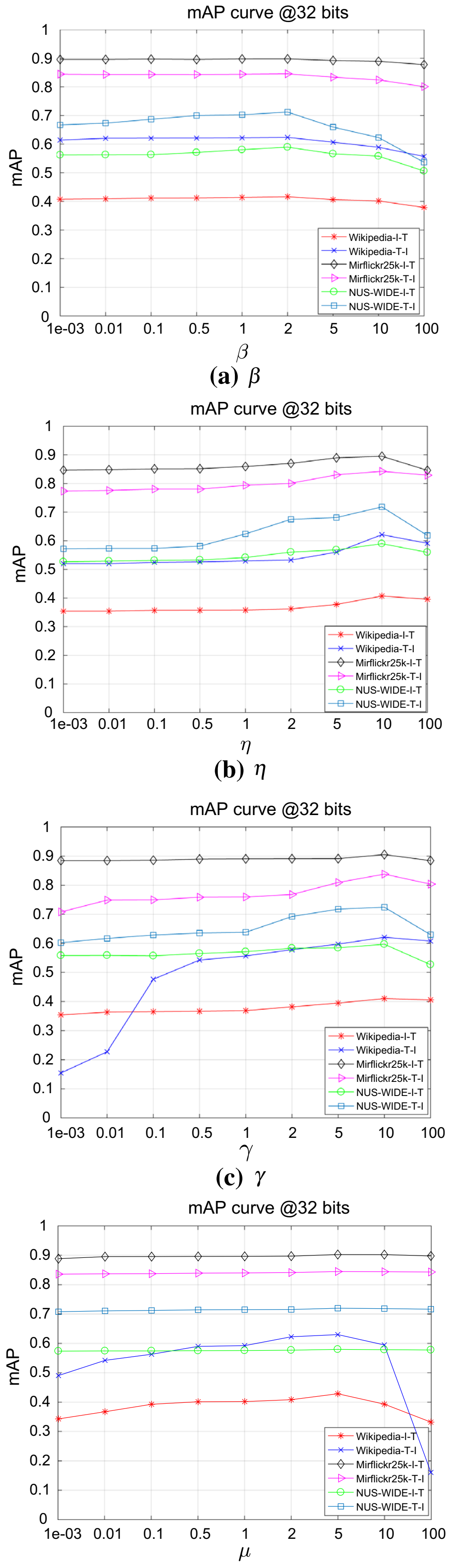


Fig. 9 Convergence curves on Wikipedia, Mirflickr25k and NUS-WIDE datasets



(a) β

(b) η

(c) γ

(d) μ

◂ **Fig. 10** The mAP results on Wikipedia, Mirflickr25k and NUS-WIDE datasets with 32 bits when the parameters β, η, γ and μ are varied from 10^{-3} to 100

Acknowledgements Weihua Ou is the corresponding author. This work was supported by the National Natural Science Foundation of China (No.61762021, 61962010, 61976107, 61876093), Natural Science Foundation of Guizhou Province (Grant No.[2017]1130, [2017]5726-32), Excellent Young Scientific and Technological Talent of Guizhou Province ([2019]-5670), the National Natural Science Foundation of Jiangsu Province under Grant BK20181393 and in the part by the China Scholarship Council under Grant 201908320072.

References

Akaho S (2006) A kernel method for canonical correlation analysis. arXiv preprint cs/0609071

Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In The 30th International Conference on Machine Learning (ICML), pages 1247–1255

Bibi R, Mehmood Z, Yousaf RM, Saba T, Sardaraz M, Rehman A (2020) Query-by-visual-search: multimodal framework for content-based image retrieval. J Ambient Intell Humaniz Comput 1–20

Chen Z-D, Li C-X, Luo X, Nie L, Zhang W, Xu X-S (2019) Scratch: A scalable discrete matrix factorization hashing framework for cross-modal retrieval. In: IEEE Transactions on Circuits and Systems for Video Technology, PP:1–1

Chua T S, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of singapore. In: ACM International Conference on Image and Video Retrieval, page 48

Deng C, Chen Z, Liu X, Gao X, Tao D (2018) Triplet-based deep hashing network for cross-modal retrieval. IEEE Trans Image Process 27(8):3893–3903

Deng C, Yang E, Liu T, Li J, Liu W, Tao D (2019) Unsupervised semantic-preserving adversarial hashing for image search. IEEE Trans Image Process 28(8):4032–4044

Ding G, Guo Y, Zhou J (2014) Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2075–2082

Ding G, Guo Y, Zhou J, Gao Y (2016) Large-scale cross-modality search via collective matrix factorization hashing. IEEE Trans Image Process 25(11):5427–5440

Gong Y, Ke Q, Isard M, Lazebnik S (2014) A multi-view embedding space for modeling internet images, tags, and their semantics. Int J Comput Vis 106(2):210–233

Huimin L, Zhang Ming X. X, Li Y, Shen H (2020) Deep fuzzy hashing network for efficient image retrieval. IEEE Trans Fuzzy Syst, PP:1–1

Huiskes MJ, Lew MS (2008) The mir flickr retrieval evaluation. In: ACM Sigmm International Conference on Multimedia information retrieval, Mir 2008. Vancouver, British Columbia, Canada October, pp 39–43

Hussain DM, Surendran D (2020) The efficient fast-response content-based image retrieval using spark and mapreduce model framework. J Ambient Intell Humaniz Comput 1–8

Jacobs DW, Daume H, Kumar A, Sharma A (2012) Generalized multiview analysis: A discriminative latent space. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pages 2160–2167

Li C, Chen Zhenduo ZP-F, Luo X, Nie L, Zhang W, Xu X-S (2018) Scratch: A scalable discrete matrix factorization hashing for cross-modal retrieval. In: Proceedings of the 26th ACM international conference on Multimedia, pages 1–9

Li C, Zhou B (2020) Fast key-frame image retrieval of intelligent city security video based on deep feature coding in high concurrent network environment. J Ambient Intell Humaniz Comput, 1–9

Lichao D, Li N, Liu W, Gao X, Tao D (2018) Self-supervised adversarial hashing networks for cross-modal retrieval. Comput Vis Pattern Recogn, p 4242–4251

Likai Qi G-J, Hua K A (2018) Learning label preserving binary codes for multimedia retrieval: a general approach. ACM Trans Mult Comput Commun Appl (TOMM) 14(1):1–23

Lu H, Li Y, Chen M, Kim H, Serikawa S (2018) Brain intelligence: go beyond artificial intelligence. Mobile Netw Appl 4(23):368–375

Lu W, Zhang X, Lu H, Li F (2020) Deep hierarchical encoding model for sentence semantic matching. J Vis Commun Image Represent, p 102794

Ma D, Liang J, Kong X, He R (2016) Frustratingly easy cross-modal hashing. In: Proceedings of the 24th ACM international conference on Multimedia, p 237–241. ACM

Ou W, Xuan R, Gou J, Zhou Q, Cao Y (2019) Semantic consistent adversarial cross-modal retrieval exploiting semantic similarity. Multimed Tools Appl 1–18

Peng Y, Huang X, Zhao Y (2017) An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. in: IEEE Transactions on Circuits and Systems for Video Technology, 1–14

Peng Y, Qi J, Yuan Y (2018) Modality-specific cross-modal similarity measurement with recurrent attention network. IEEE Trans Image Process 27(11):5585–5599

Rasiwasia N, Pereira JC, Coviello E, Doyle G, Lanckriet GRG, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: International Conference on Multimedia, 251–260

Schönemann PH (1966) A generalized solution of the orthogonal procrustes problem. Psychometrika 31(1):1–10

Singh AP, Gordon GJ (2008) Relational learning via collective matrix factorization. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 650–658

Song J, Yang Y, Yang Y, Huang Z, Shen HT (2013) Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp 785–796

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

Wang D, Gao X, Wang X, He L (2015) Semantic topic multimodal hashing for cross-media retrieval. In: the International Joint Conference on Artificial Intelligence(IJCAI), pages 2291–2297

Wang D, Lu H (2013) On-line learning parts-based representation via incremental orthogonal projective non-negative matrix factorization. Signal Process 93(6):1608–1623

Wang D, Lu H, Yang M-H (2016) Robust visual tracking via least soft-threshold squares. IEEE Trans Circuits Syst Video Technol 26(9):1709–1721

Wang W, Livescu K (2015) Large-scale approximate kernel canonical correlation analysis. arXiv preprint arXiv:1511.04773

Xu X, He L, Shimada A, Taniguchi RI, Lu H (2017) Learning unified binary codes for cross-modal retrieval via latent semantic hashing. Neurocomputing 213:191–203

Xu X, Lu H, Song J, Yang Y, Shen HT, Li X (2019) Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. IEEE Trans Cybern (**in press**)

Xu XS (2017) Dictionary learning based hashing for cross-modal retrieval. In: Proceedings of the 24th ACM international conference on Multimedia, pp 177–181

Yang Y, Zhuang Y-T, Wu F, Pan Y-H (2008) Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Trans Multimed 10(3):437–446

Yao T, Han Y, Tao WR, Kong X, Yan L, Fu H, Tian Q (2019) Efficient discrete supervised hashing for large-scale cross-modal retrieval. arXiv preprint arXiv:1905.01304

Yao T, Kong K, Fu H, Tian Q, (2019) Discrete semantic alignment hashing for cross-media retrieval. IEEE Trans Cybern 99:1–12

Yaotao Zhang, Z, Yan L, Yue J, Tian Q (2019) Discrete robust supervised hashing for cross-modal retrieval. IEEE Access 7:39806–39814

Zhang D, Li W-J (2014) Large-scale supervised multimodal hashing with semantic correlation maximization. AAAI 2177–2183

Zhang Y, Lu W, Ou W, Zhang G, Zhang X, Cheng J, Zhang W (2020) Chinese medical question answer selection via hybrid models based on cnn and gru. Multimed Tools Appl 1–26

Zhou J, Ding G, Guo Y (2014) Latent semantic sparse hashing for cross-modal similarity search. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 415–424

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.