Applied Intelligence (2022) 52:17170–17184
https://doi.org/10.1007/s10489-022-03367-w

Adaptive multi-modal fusion hashing via Hadamard matrix

Jun Yu[1] · Donglin Zhang[2] · Zhenqiu Shu[3] · Feng Chen[4]

Accepted: 8 February 2022 / Published online: 31 March 2022

Abstract
Hashing plays an important role in information retrieval, due to its low storage and high speed of processing. As an effective multi-modal representation learning method, multi-modal hashing has received particular attention. Most of the existing multi-modal hashing methods adopt the fixed weighting factors to fuse multiple modalities for any query data, which cannot capture the variation among different queries. Besides, there are too much hyper-parameters in their models while it is time-consuming and labor-intensive to determine the proper parameters. The limitations may significantly hinder their promotion in practical applications. In this paper, we propose a simple, yet effective method that is inspired by the Hadamard matrix. On the one hand, our proposed method that involves a very few hyper-parameters is flexible. On the other hand, the complementary information between multi-modal data and the semantic discrimination information are preserved well in the hash codes. Extensive experimental results on four benchmark datasets show that the proposed framework is effective and achieves superior performance compared to state-of-the-art methods.

Keywords Multi-modal hashing · Hadamard matrix · Hash centers · Adaptively dynamic weights

1 Introduction

Multi-modal scenes are very popular with the development of sensor technology. As an effective technique to deal with the challenges posed by the explosive growth of multi-modal data, multi-modal hashing has attracted increasing attention in information retrieval and related areas [25]. Among the techniques available in the literature, a simple way is extending uni-modal hashing to the multi-modal situation. Concretely, multiple modality features are concatenated and regarded as the input of uni-modal hashing methods. Such as Locality Sensitive Hashing (LSH) [3], Iterative Quantization (ITQ) [4], Discrete Locally Linear Embedding (DLLE) [6], Hadamard Codebook based Online Hashing (HCOH), Supervised Discrete Hashing (SDH)[19] etc. Nevertheless, these hashing methods mainly focus on uni-modal hashing that implements a query process within a single modality, such an extension may cause "Curse of Dimensionality". To handle this problem, multi-modal hashing methods combine multiple modalities to comprehensively represent data in multimedia retrieval applications. Some learning methods [14, 20, 21, 27] have been developed. Multiple Feature Hashing (MFH) [21] explores the local structure of individual modality and fuses them in a joint framework. Multiple Kernel Learning (MKH) [14] fuses multiple modalities with the aid of an optimised linear-combination. Multi-view Latent Hashing (MVLH) [20] aims to find a unified kernel feature space where the weights of different modality are adaptively learned. Multiview discrete hashing (MVDH) [27] jointly performs a matrix factorization and spectral clustering to

✉ Jun Yu
yujun@zzuli.edu.cn

Donglin Zhang
dlinzzhang@gmail.com

Zhenqiu Shu
shuzhenqiu@163.com

Feng Chen
571435345@qq.com

1 The College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, China

2 The School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China

3 Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

4 School of Computer Science and Technology, Anhui University of Technology, Ma'anshan, China

learn compact hash codes. In MVDH, the weight of each modality is adaptively learned to reflect the importance of the corresponding modality. However, these methods do not sufficiently consider the differences between different data, applying the fixed weights to encode query data. To tackle this issue, some online adaptive hashing methods [17, 36] are proposed. An example is Online Multi-modal Hashing with Dynamic Query-adaption (OMH-DQ) [17] where an online parameter-free mode is designed to adaptively learn the hash codes for the dynamic queries. These approaches have achieved promising performance in many applications. Unfortunately, the adjacent graph construction for each mod-ality takes significant storage and computing cost, which is not scalable to large-scale multimedia data. Furthermore, the additional hyper-parameters are introduced to balance the regularization terms in their models and it is tough to obtain the optimal parameters.

Based on the above observations, in this paper, we propose a simple yet very effective multi-modal hashing to overcome the challenges. Inspired by some recent works [8, 10, 31] where Hadamard matrix has been verified to be effective in the hash learning, we employ the properties of Hadamard matrix to generate discriminative target codes for multimodal data. The proposed method induces the samples with the same label to be close to their common target codes in the hash function learning stage. In the hash encoding stage, we adopt an adaptively self-weighting scheme to capture the dynamic variation information between query data. Figure 1 illustrates the flowchart of the proposed framework. The advantages of our method are summarized as follows

- We introduce a Hadamard matrix into the multi-modal learning framework to guide the hash function learning. Our model enables the discriminative semantic information to be preserved in the hash codes, although our model is concise.
- An adaptive and self-weighted hash encoding module can accurately perceive the modality variations of dynamic queries, which enhance the robustness of our model in more complex data scenarios.
- A comparative evaluation of our proposed method with state-of-the-art methods on four available datasets shows that our proposed method boosts the retrieval performance.

Structurally, the rest of this paper is organized as follows. In Section 2, we review related works. The proposed model is described in Section 3. In Section 4, we analyse the

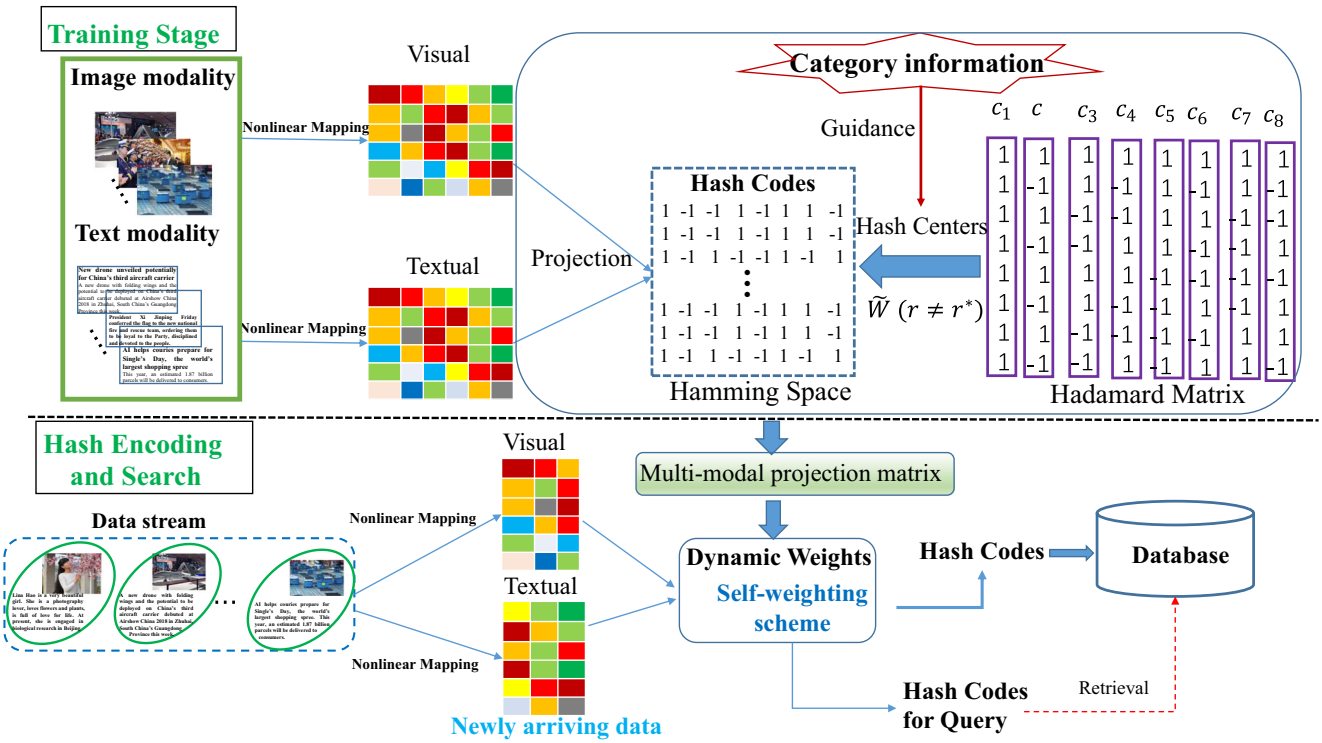


Fig. 1 The overview of our proposed method. The proposed framework contains two parts: Training stage and Hash encoding stage. The multi-modal data is collaboratively projected into the common Hamming space where samples of the same category converge to their common hash center point generated by a Hadamard matrix in the offline training stage. Based on the learned projection matrices in the offline training stage, we adopt an adaptive weighing scheme to obtain hash codes in order to capture the variation of different samples in the hash encoding process and the hash codes of sample are archived in database

experimental results. The conclusions of the paper is drawn in Section 5.

2 Related work

As mentioned previously, many hashing methods have been proposed to handle the multi-modal data, which can be grouped into Cross-modal hashing and Multi-modal hashing. The cross-modal hashing [5, 9, 13, 24, 30, 33] learns a hash function for each modality to encode the hash codes of the corresponding modality. Discrete Cross-modal Hashing (DCH) [28] is a joint optimization problem to learn the modality-specific hash functions and the unified binary codes simultaneously. DCH only learns the discriminative binary codes from the perspective of classification, but the semantic label information is not utilized adequately. Multimodal Discriminative Binary Embedding (MDBE) [23] formulates the hash function learning in terms of classification, exploiting the label information to discover the shared structures inside heterogeneous data. Nevertheless, the linear hash function is inferior to nonlinear functions since it is not able to capture the non-linear structure. Lin Z. et al. [12] proposed Semantics-Preserving Hashing(SePH) in which the kernel logistic regression with a sampling strategy is employed to learn the nonlinear projections from the original space to Hamming space. SePH transforms semantic affinities into a probability distribution and approximates it in Hamming space via minimizing the Kullback-Leibler divergence. Discrete Latent Factor model based cross-modal Hashing (DLFH) [7] directly learns the binary hash codes without continuous relaxation. DLFH achieves better performance, but the training cost is high. Enhanced Discrete Multi-modal Hashing (EDMH) [1] learns binary codes and hashing functions from the pairwise similarity matrix of data. Since a fast iterative learning algorithm is developed, EDMH runs much faster than most of cross-modal hashing method. However, the pairwise similarity matrix increases storage consumption.

Different from cross-modal hashing, multi-modal hashing [15–17, 20, 21, 34–36] learns a unified hash function for paired multi-modal data. In other words, the paired multi-modal information is fused in discrete Hamming space. Compared with cross-modal hashing, it can better represent a target object since the complementary information between multiple modalities is fully mined. Song J. et al. [21] proposed Multiple Feature Hashing (MFH) that not only preserves the local structure information of each individual mod-ality but also globally considers the geometric structure of all modalities to learn a group of hash functions. MFH is very stable to well characterize the visual content. However, MFH is not easy to be extended to large-scale data scenarios as the cost of affinity matrix construction is proportional to the square of the number of samples. Multi-view Latent Hashing (MVLH) [20] explores the latent factors shared by multiple views in a unified kernel feature space to learn the binary codes. MVLH reveals the nonlinear data structures of different views, but the bit-by-bit discrete optimization consumes considerable computation time. Flexible Multi-modal Hashing (FMH) [36] learns multiple modality-specific hash codes and the combination of modality features is flexibly generated for the newly coming queries. The concatenated combination raises the redundant information of feature space. Efficient Parameter-Free Adaptive Multi-modal Hashing (EPAMH) [35] is proposed to automatically and ada-ptively determine the modality combination weights, but it does not utilize the semantic information to enhance the discriminative ability of model. Semantic-driven Interpretable Deep Multi-modal Hashing (SIDMH) [15] employs a deep hashing architecture driven by semantic categories to generates hash codes. Online Multi-modal Hashing with Dynamic Query-adaption (OMH-DQ) [17] method is designed to preserve the pair-wise semantic similarity and the complementary information between multi-modal features by hash codes. Fast Discrete Collaborative Multi-Modal Hashing (FDCMH) [34] preserves the semantic correlations described by pair-wise similarity matrix into hash codes in a discrete hashing learning framework. However, the pair-wise semantic affinity matrix calculation increases the algorithm complexity. Besides, the above multi-modal hashing methods introduce additional parameters whose values need to be manually adjusted. This indicates that more time is needed to acquire the optimal parameters.

In this paper, we propose a novel multi-modal hashing algorithm that utilizes the advantages of the Hadamard matrix and the category information of data to guide the hash learning. In the hash encoding process, our proposed model adopts the adaptive weighting way to capture the dynamic difference between multi-modal data. The proposed method is implemented conveniently since it hardly needs to be manually set proper parameter in the learning process.

3 The proposed method

3.1 Model formulation

Assume that the training dataset $\chi = \{o_1, o_2, ..., o_n\}$ is comprised of n multi-modal instances and a sequence set $Y = \{y_1, y_2, ..., y_n\}$, where y_i denotes the category information of o_i. Each instance consists of paired M

modalities. The m-th modality feature matrix is denoted as $X^{(m)} = \left[x_1^{(m)}, ..., x_n^{(m)}\right] \in R^{d_m \times n}$, where d_m is the dimensionality of the m-th modality. Our method aims to learn the discriminative hash code $B \in \{-1, 1\}^{r \times n}$ to represent multi-modal instances in Hamming space, where r is the length of the to-be-learnt hash codes. We pre-define a set of points $C = \{c_1, c_2, ..., c_s\} \in R^{r^* \times s}$, where s and r^* are the number of categories and the dimension of hash centers, respectively. c_i is the hash center of i-th category data. $r^* = r$ indicates the dimension of hash center is same as that of the Hamming space. We encourage the data points belonging to the same class to be close to their common hash center and those with different semantic categories to be associated with different hash centers respectively. Intuitively, the pre-defined hash centers should conform to the following requirement: A sufficient mutual distance between the centers should be ensured so that samples from different classes are separated well in the Hamming space. The concept of a valid hash center set is summarized in Definition 1 [10, 31].

Definition 1 Hash center set $C = \{c_i\}_{i=1}^s \subset \{0, 1\}^{r^*}$ in the r^*-dimensional Hamming space satisfies that the average pairwise Hamming distance is greater than or equal to $r^*/2$, i.e.

$$\frac{1}{V}\sum_{i \neq j}^{s} D_H(c_i, c_j) \geq \frac{r^*}{2} \tag{1}$$

where V is the number of different combinations about c_i and c_j; s is the size of C; D_H denotes the Hamming distance.

The recent work [8] is proposed to simplify the learning-based supervised discrete hashing (SDH) by a "Hadamard matrix". [10] extends the Hadamard-based SDH to online scenario where data is received in a streaming manner. [31] learns a non-linear hash function via reducing the quantization error between the to-be-learnt hash codes and the corresponding column of Hadamard matrix. Inspired by [8, 10, 31], we introduce the Hadamard matrix to guide the multi-modal hashing learning. It is noted that [8, 10, 31] are limited to uni-modal hashing scenarios with single modality data, which is not suitable to handle multi-modal hash learning problem. The purpose of our method is to effectively fuse multi-modal data information and learn high-quality and discriminative hash codes by employing the characteristics of Hadamard matrix. A Hadamard matrix constructed by Sylve-ster method [22] has the following properties:

- It is an r^*-order ($r^* = 2^t, t = 1, 2...$) squared matrix whose elements are either +1 or -1. The dimension of the generated Hadamard matrix is

$$r^* = \min\{l | l = 2^t, l \geq r, l \geq k, t = 1, 2, ...\} \tag{2}$$

- Its columns are pair-wise orthogonal, which ensures the Hamming distance between any two column vectors is $r^*/2$.

Thus, the above construction conforms to Definition 1 and each column of Hadamard matrix can serve as a hash center.

Referring to (2), there may be cases when the output code length r does not satisfy: $r = r^*$. To mitigate this problem and ensure the dimension of centers is consistent with the output codes, Local Sensitive Hashing(LSH) is adopted to transform the hash centers generated by Hadamard matrix into r-dimension Hamming space.

$$\tilde{C} = sign(\tilde{W}^T C) \tag{3}$$

where $\tilde{W} = \{\tilde{w}_i\}_{i=1}^r \in R^{r^* \times r}$ is sampled from the standard Gaussian distribution. $sign(x)$ denotes the sign function that returns '+1' if the variable x greater than or equal '0', and '-1' otherwise. $\tilde{C} = \{\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_s\} \in R^{r^* \times s}$ preserves the main properties of the original matrix C and complies with the requirement of minimal average Hamming distance among columns. The detailed theory is developed in [10]. Thus, we have

$$\bar{C} = \begin{cases} C & r = r^* \\ \tilde{C} & r \neq r^* \end{cases} \tag{4}$$

As described above, $\bar{C} = \{\bar{c}_1, \bar{c}_2, ..., \bar{c}_s\}$ is a center set, where $\bar{c}_i$ relates to i-th category. For single-label data, each multi-modal instance is assigned to one hash center and those samples with the same category share a common center. All samples are encouraged to be close to corresponding hash center. The target hash codes for the training set χ are denoted as $H^* = \{c_1^*, c_2^*, ..., c_n^*\} \in R^{r \times n}$, where $c_i^* \in \bar{C}$ is the hash center of multi-modal instance o_i. For multi-label data that is classified into two or more categories, the target hash codes are represented by the centroid of the multiple centers, each of which relates to a single category. For example, assume one multi-modal instance o_i that has been classified into four categories y_i, y_j, y_k and y_g. The target codes c_i^* of multi-modal instance o_i become the centroid c of the corresponding four hash centers $\bar{c}_i$, $\bar{c}_j$, $\bar{c}_k$ and $\bar{c}_g$, where c is calculated by voting at the same bit and taking the dominative value. As shown in Fig. 2, we first vote for '+1' or '-1' based on the maximum number of bits and vote for 0 if the number of '+1' is equal to that of '-1'. Then all '0' will be adjusted to '+1' or '-1' by sampling from Bernoulli distribution $Bern(0.5)$. Therefore, the above strategy can well extend our proposed framework to the multi-label data.

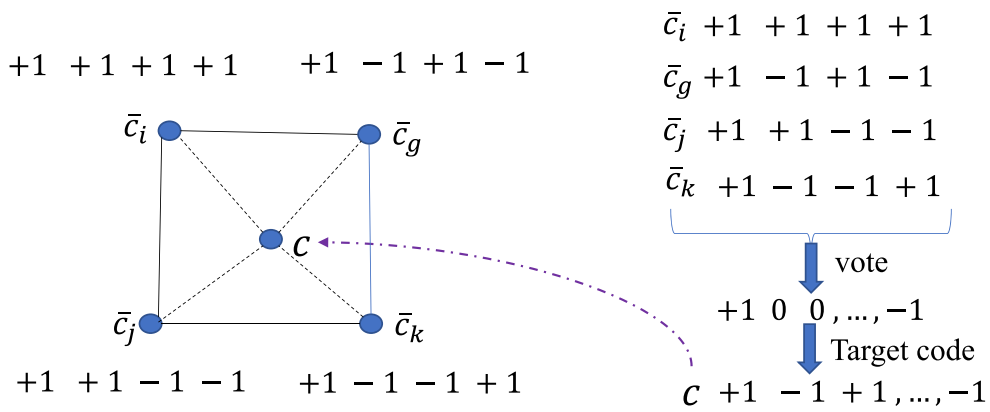


Fig. 2 A toy example of Hash center for multi-label data

3.2 Training stage

Through the above process, we obtain the target hash representation $H^* \in R^{r\times n}$ for the training set χ in the Hamming space. For the m-th modality $X^{(m)} = \left[x_1^{(m)}, ..., x_n^{(m)}\right] \in R^{d_m\times n}$, we calculate a nonlinearly transformed representation $\phi\left(x_i^{(m)}\right) = \left[exp\left(\frac{\|x_i^{(m)}-a_1^{(m)}\|_F^2}{2\sigma_m^2}\right), ..., exp\left(\frac{\|x_i^{(m)}-a_p^{(m)}\|_F^2}{2\sigma_m^2}\right)\right]$, where σ_m denotes the Gaussian kernel parameter and $\left\{a_j^{(m)}\right\}_{j=1}^{p}$ are p anchors that are randomly selected from the m-th modality of the training data. $\phi(X^{(m)}) = \left[\phi\left(x_1^{(m)}\right), ...\phi\left(x_n^{(m)}\right)\right] \in R^{p\times n}$ preserves the intra-modal correlation. The time complexity of the phase is $\mathcal{O}(Mnp)$, which is linear to the size of the training set.

The heterogenous modalities are projected into a common Hamming space. In this space, data points are encouraged to migrate towards their target hash codes. Thus, we have the following

$$\min_{W^{(m)}} \sum_{m=1}^{M} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F \tag{5}$$

where $W^{(m)} \in R^{r\times p}$ is the projection matrix of the m-th modality; H^* is the target representation matrix for the training samples; $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

In multimedia retrieval, there may potentially be discrepancy between the heterogeneous modalities. Accordingly, it is necessary to gauge the importance of different modalities so as to learn an effective and discriminative hash function. To address the problem, we transform (5) to its equivalent form (see Proof 1):

$$\begin{aligned} \min_{\mu^{(m)}, W^{(m)}} & \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2 \\ s.t. & \sum_{m=1}^{M} \mu^{(m)} = 1 \end{aligned} \tag{6}$$

As formulated in (6), $\frac{1}{\mu^{(m)}}$ can be considered as the function of $\mu^{(m)}$. The more discriminative the m-th modality is, the smaller $\|H^* - W^{(m)}\phi(X^{(m)})\|_F^2$ and $\mu^{(m)}$ should be, and vice versa.

Proof Equation (5) is equivalent to (6). □

According to the Cauchy-Schwarz inequality, the following (7) holds.

$$\begin{aligned} & \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2 \\ \Leftrightarrow & \left(\sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2\right) \\ & \times \left(\sum_{m=1}^{M} \mu^{(m)}\right) \\ \geq & \left(\sum_{m=1}^{M} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F\right)^2 \end{aligned} \tag{7}$$

Thus, we can obtain

$$\begin{aligned} & \left(\sum_{m=1}^{M} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F\right)^2 \\ = & \min_{\mu^{(m)}} \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2 \end{aligned} \tag{8}$$

then

$$\begin{aligned} & \min_{W^{(m)}} \sum_{m=1}^{M} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F \\ \Leftrightarrow & \min_{W^{(m)}} \left(\sum_{m=1}^{M} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F\right)^2 \\ \Leftrightarrow & \min_{\mu^{(m)}, W^{(m)}} \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2 \end{aligned} \tag{9}$$

To avoid over-fitting, a regularization term is added to (6). The overall learning framework becomes

$$\begin{aligned} \min_{\mu^{(m)}, W^{(m)}} & \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - sign(W^{(m)}\phi(X^{(m)}))\|_F^2 \\ & + \delta \sum_{m=1}^{M} \|W^{(m)}\|_F^2 \\ s.t. & \sum_{m=1}^{M} \mu^{(m)} = 1 \end{aligned} \tag{10}$$

where δ is a penalty parameter. We relax the objective function into (11) without the sign function to make it tractable computationally, since it is difficult to optimize (10) directly.

$$\begin{aligned} \min_{\mu^{(m)}, W^{(m)}} & \sum_{m=1}^{M} \frac{1}{\mu^{(m)}} \|H^* - W^{(m)}\phi(X^{(m)})\|_F^2 + \delta \sum_{m=1}^{M} \|W^{(m)}\|_F^2 \\ s.t. & \sum_{m=1}^{M} \mu^{(m)} = 1 \end{aligned} \tag{11}$$

We adopt the alternative optimization method to solve the relaxed problem (11).

- Update $W^{(m)}$ with other variables fixed. We set the derivative of the objective function (11) with respect to $W^{(m)}(m = 1, 2....)$ to zero and then the solution is obtained as follows

$$W^{(m)} = \frac{1}{\mu^{(m)}} H^* \phi^T(X^{(m)}) \left(\frac{1}{\mu^{(m)}} \phi(X^{(m)}) \phi^T(X^{(m)}) + \delta I \right)^{-1} \tag{12}$$

- Update $\mu^{(m)}$ with other variables fixed. Gathering the terms relating to $\mu^{(m)}$, we gain the subproblem

$$\begin{aligned} \min_{\mu^{(m)} \geq 0} & \sum_{m=1}^{M} \frac{(G^{(m)})^2}{\mu^{(m)}} \\ s.t. & \sum_{m=1}^{M} \mu^{(m)} = 1 \end{aligned} \tag{13}$$

where $G^{(m)} = \|H^* - W^{(m)}\phi(X^{(m)})\|_F$. According to the Cauchy-Schwarz inequality, we obtain the optimal $\mu^{(m)}$ as follows

$$\mu^{(m)} = \frac{G^{(m)}}{\sum_{m=1}^{M} G^{(m)}} \tag{14}$$

After multiple iterations, we obtain the optimal $W^{(m)}$ and $\mu^{(m)}$. The hash function $f = sign(\sum_{m=1}^{M} \frac{1}{\mu^{(m)}} W^{(m)} \phi(X_q^{(m)}))$ that fuses multiple modal information can be used to generate the hash codes for multi-modal instances.

3.3 Hash encoding with dynamic weights

The learned weights $\mu^{(m)}$ are fixed in hash function f. Unfortunately, the fixed weights can not capture the modality variations of dynamic data in the process of hash encoding. Thus, they should be adjusted dynamically for different multi-modal instances. Motivated by this intuition, we adopt an online self-weighting scheme for newly arriving multi-modal data to obtain more accurate hash codes. In the encoding stage, we assume that data appear in the manner of the data stream. The adaptive online hash encoding process is given as follows

$$\begin{aligned} \min_{B_q, \mu_q^{(m)}} & \sum_{m=1}^{M} \frac{1}{\mu_q^{(m)}} \|B_q - W^{(m)}\phi(X_q^{(m)})\|_F^2 \\ s.t. & \sum_{m=1}^{M} \mu^{(m)} = 1, B_q \in \{-1, 1\}^{r \times n_q} \end{aligned} \tag{15}$$

where B_q and n_q are the hash codes and the number of the new instances, respectively. $\phi(X_q^{(m)})$ is the nonlinearly feature representation of the m-th modality of the newly coming instances.

Algorithm 1 Adaptive multi-modal fusion hashing.

Input: Training set $X^{(m)} = [x_1^{(m)}, ..., x_n^{(m)}] \in R^{d_m \times n}$, $(m = 1, ..., M)$ with category information.

Output: $W^{(m)}(m = 1, ..., M)$, $B_q(q = 1, ...T)$

% step-1:projection learning stage

1. Generate target codes H^*
2. Calculate the transformed representation $\phi(X^{(m)}) = [\phi(x_1^{(m)}), ...\phi(x_n^{(m)})] \in R^{p \times n}$
3. Initialize $W^{(m)}(m = 1, ..., M)$ and $\mu^{(m)}(m = 1, ..., M)$

Repeat

4. Update $W^{(m)}(m = 1, ..., M)$ according to (12);
5. Update $\mu^{(m)}(m = 1, ..., M)$ according to (13);

Until convergence

% step-2:Hash encoding stage

for q =1,...T do

6. Receive newly arriving X_q and calculate the non-linear representation $\phi(X_q^{(m)})$.

Repeat

7. Update $\mu^{(m)}(m = 1, ..., M)$ according to (16);
8. Update B_q according to (17);

Until convergence

end for

The problem in (15) is solved by updating the following variables alternately.

Update $\mu_q^{(m)}$ with fixed B_q. The optimal solution of $\mu_q^{(m)}$ is obtained as follows

$$\mu_q^{(m)} = \frac{G_q^{(m)}}{\sum_{m=1}^{M} G_q^{(m)}} \tag{16}$$

where $G_q^{(m)} = \|B_q - W^{(m)}\phi(X_q^{(m)})\|_F$.

Update B_q with fixed $\mu_q^{(m)}$. We can get a closed solution:

$$B_q = sign\left(\sum_{m=1}^{M} \frac{1}{\mu_q^{(m)}} W^{(m)} \phi(X_q^{(m)}) \right) \tag{17}$$

The above optimization process is conducted iteratively and the optimal B_q is viewed as the final output codes. The output hash codes of the newly arriving data will be archived in the database. The proposed Adaptive Multi-modal Fusion Hashing (AMFH) is summarized in Algorithm 1. For the modality missing problem in which the partial modalities is unknown, the corresponding weighting factor is set to zero. The main difference between our method and the existing Hadamard matrix-based hashing methods like [8, 10, 11] is summarized as follows. The previous methods only model single modality data to realize uni-modal retrieval, such as image retrieval. Differently, the proposed method in this

paper fuses multiple modalities into the hash learning in an adaptive weighting manner and provides the solution idea to generate the target hash codes for multi-modal instance data with multiple categories. In the experiment section, we take image modality and text modality as an example to present some discussions.

4 Experiment

In this section, we perform multi-modal retrieval experiments on four widely-used multi-modal datasets to verify the effectiveness of our proposed method. The extended cross-modal retrieval experiments are also carried out to investigate the performance of the proposed method. Figure 3 shows two possible data scenarios. For cross-modal retrieval scenario, the paired modalities relationship is not met strictly for multi-modal instances and the query data is single modality. Our experiments are executed on a Windows 10 platform based desktop machine with 12GB memory and 4-core 3.6GHz CPU.

4.1 Datasets

WiKi [18] is a multi-modal single-label dataset which consists of 2,866 multimedia documents. The dataset is divided into 10 categories. We assign one hash center for each category. Each image and text are represented by 128-dimensional SIFT histogram vector and 10-dimensional feature vector generated by latent Dirichlet allocation, respectively. A random subset with 2,173 multimedia documents is used as the offline training set and the remaining 963 samples as the query set.

Pascal VOC 2007 [26] contains 9,963 multi-modal instances of 20 categories. Each multi-modal instance is composed of one image and 399 tags associated with the image. In this dataset, we employ the 4,096-dimensional CNN feature and the 798 dimensional tag ranking feature to represent visual and textual object, respectively. A random subset with 2,000 instances is provided to compose the offline training set. The remaining data is divided into 963 and 7,000 instances that are used as query set and retrieval set, respectively.

NUS-WIDE [2] is comprised of 26,9648 multi-modal instances of 81 concepts. In our experiments, we only keep 18,6577 instances of the top ten most frequent concepts. The image modality is represented by a 500 dimensional bag- of-visual words and the 1000 dimensional tag occurrence vector is employed as text modality feature. A random subset with 1,866 instances is viewed as the query set and the remaining 18,4711 instances for retrieval set. The randomly selected 5,000 instances from the retrieval set are used to compose the offline training set.

IAPR-TC12 [2] consists of 20,000 multi-modal instances of 255 categories. Each instance is an image-text pair where text includes one or multiple tags and associates with image. In our experiments, we randomly take 2,000 multi-modal instances to form the query set. The remaining 18,000 instances are used as the retrieval set. We employ 28% of instance data in the retrieval set as the training set. The image modality and the text modality are represented by the 512-dimensional GIST feature vector and the 2,912-dimensional bag-of-words vector, respectively.

Pascal VOC 2007, NUS-WIDE and IAPR-TC12 are three multi-label datasets. According to Section 3.1, we first obtain a hash center set in which each element is assigned for the hash center of a single category, and then calculate the centroid of multiple hash centers for the multi-label data as their target codes.

4.2 Experimental settings

Our proposed ADFH is compared with some state-of-the-art hashing methods in multi-modal retrieval field. These baselines can be divided into two groups. (1) Multi-modal hashing methods, such as MFH [21], MVLH [20], FOMH

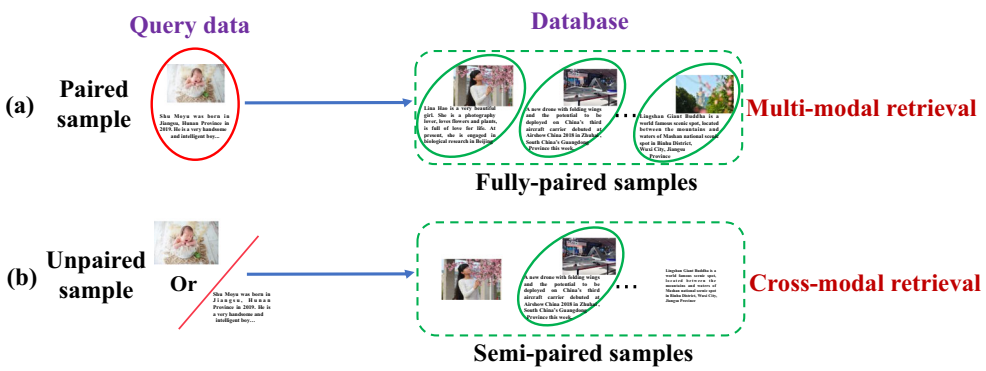


Fig. 3 Two data scenarios in hash encoding process

[16], OMH-DQ [17], EPAMH [35], FDCMH [34] and SIDMH [15]; (2) Uni-modal hashing methods, such as ITQ [4], LSH [3], DLLE [6], HCOH [11]. Since the uni-modal methods can not deal with multiple modalities directly, we concatenate multiple modalities for fair comparison. EPAMH is only suitable for this case where the dimension of the original multi-modal feature space is larger the length of hash codes. EPAMH is not compared on WiKi dataset since its textual dimension is less than 16. It is noted that the text features and image features described in Section 4.1 are imported into the deep module of the SIDMH framework to replace its feature input. In addition, we also compare our method with the following six supervised cross-modal hashing methods: MDBE [23], SePH [12], DCH [28], DLFH [7], KDLFH [7], E-DMH [1]. We adjust the parameters of each baseline in the candidate range that is given in the original paper and the best results are reported in this paper. The performance of all method is evaluated by Mean Average Precision (mAP) [29, 32]. For a given query q, the Average Precision (AP) is defined as follows:

$$AP(q) = \frac{1}{l_q} \sum_{m=1}^{R} P_q(m)\delta_q(m) \tag{18}$$

where $P_q(m)$ denotes the accuracy of the top m retrieval results; $\delta_q(m) = 1$ if the m-th position is the true neighbour of the query q, and otherwise $\delta_q(m) = 0$; l_q is the correct statistics of top R retrieval results. The mAP is defined as the mean of the AP of all queries.

4.3 Retrieval accuracy comparison

1) *Results on WiKi:* The experimental results on WiKi are reported in Table 1. We can clearly observe that our method consistently outperforms all baselines when the code length varies from 16 bits to 128 bits. With the code length increases, the performance of our method improves slightly. The results demonstrates that our method is not sensitive to the code length and is able to achieve satisfactory performance even with short codes. AMFH is superior to the uni-modal hashing methods with different hash code lengths. The possible reason for better performance of AMFH is that it can reduce the redundant information among multiple mod-alites. Compared with OMH-DQ, our method achieves an average improvement of 23% on WiKi dataset. The promising comparison results show that our method which assigns the semantic hash centers for different category data via Had-amard matrix can promote the discriminative hash function learning. AMFH also achieves much higher mAP scores than other multi-modal hashing methods, for the reason that the adaptive online self-weighting strategy can improve the quality of hash codes. The comparison results show that the hash learning framework based on Hadamard matrix is effective for single-label data.

Table 1 mAP Comparison of multi-modal retrieval for different bits on WiKi

Methods	WiKi			
	16	32	64	128
ITQ [4]	0.5122	0.5359	0.5490	0.5532
LSH [3]	0.4306	0.4712	0.5085	0.5276
DLLE [6]	0.5234	0.5330	0.5466	0.5506
HCOH [11]	0.5450	0.5474	0.5494	0.5490
MFH [21]	0.4630	0.5040	0.5455	0.5569
MVLH [20]	0.3027	0.3166	0.3000	0.3045
FOMH [16]	0.4408	0.4813	0.5092	0.5149
OMH-DQ [17]	0.4117	0.4393	0.4556	0.4319
EPAMH [35]	–	–	–	–
FDCMH [34]	0.5986	0.6392	0.6409	0.6531
SIDMH [15]	0.5666	0.6219	0.6307	0.6416
Ours	**0.6580**	**0.6674**	**0.6677**	**0.6752**

Bold entries signify the best result

2) *Results on Pascal VOC 2007:* The experimental results on Pascal VOC 2007 are presented in Table 2. Our proposed AMFH consistently obtains the best average mAP score than baselines. As shown in Table 2, the performance of our method is inferior to FDCMH and SIDMH when the code length is set to 16 bits, while our method achieves an average improvement of 3% than the best baseline as the hash code length increases from 32 bits to 128 bits. When the code length is greater than 32 bits, the code length has little impact on performance. That means that hash codes generated by our AMFH have more strong semantic discrimination capability if the hash code length is larger than the number of categories. On Pascal VOC 2007 dataset, our method outperforms OMH-DQ by an average improvement of 13% in terms of mAP result. The comparison results demonstrate that AFMH can preserve the discriminative information of inter-class for multi-modal data into the to-be-learnt hash function.
3) *Results on NUS-WIDE:* The mAP scores of all compared methods on NUS-WIDE are shown in Table 3, from which we can see that our AMFH achieves consistent improvements over all baselines. FDCMH and OMH-DQ are two classic asymmetric similarity matrix factorization models. OMH-DQ preserves the high-level semantic correlations into the hash codes via the projection learning, while FDCMH uses matrix factorization way to learn optimal binary codes. Compared with OMH-DQ and FDCMH, our method achieves an average improvement of 7% and 3.5%, respectively on NUS-WIDE. SIDMH is a deep model

Table 2 mAP Comparison of multi-modal retrieval for different bits on Pascal VOC 2007

Methods	Pascal VOC 2007			
	16	32	64	128
ITQ [4]	0.7586	0.7975	0.8053	0.8061
LSH [3]	0.4402	0.5591	0.6628	0.7262
DLLE [6]	0.7629	0.8068	0.8131	0.8193
HCOH [11]	0.2436	0.6050	0.6070	0.6072
MFH [21]	0.5364	0.6376	0.6941	0.7216
MVLH [20]	0.5469	0.6324	0.6995	0.7203
FOMH [16]	0.4436	0.5676	0.6504	0.7343
OMH-DQ [17]	0.5673	0.7040	0.8096	0.8542
EPAMH [35]	0.7704	0.8193	0.8493	0.8738
FDCMH [34]	**0.8305**	0.8517	0.8727	0.8921
SIDMH [15]	0.8138	0.8346	0.8511	0.8634
Ours	0.7715	**0.9001**	**0.9009**	**0.9016**

Bold entries signify the best result

based multi-modal hash learning architecture. Our method performs better than SIDMH, which further indicates the pre-defined hash centers based Hadamard matrix are effective to learn the discriminative hash functions for large-scale multi-modal data retrieval.

4) *Results on IAPR-TC12:* Table 4 shows the experimental results on IAPR-TC12. We can find that the accuracy of our proposed AMFH is competitive compared with baselines. Specifically, the average score of AMFH is higher than FDCMH by 1.8%. SIDMH utilizes class label and a hierarchical model to guide the generation of hash codes. Although the complementary information between multi-modal data can be well mined in SIDMH, the discrete solution in optimization is only the approximation of the optimal solution. Our method obtains better experimental results than SIDMH. The comparison results imply that the customized hash center in this paper is very discriminative, so that the different category data can be effectively separated in the Hamming space. The experimental results on Pascal VOC 2007, NUS-WIDE and IAPR-TC12 indicates that our proposed method which assigns the hash centers based Hadamard matrix for different category multi-modal samples can generate effective hash codes in multi-modal applications with multi-label data.

5) *Comparison experiments for Cross-modal retrieval:* In this section, we further explore the performance of our model for cross-modal retrieval scenario, i.e., given a modality as query to retrieve another modality data from the databa-se. As illustrated in Fig. 3b, the partial modality of some instances is missing. We need to encode these unpaired instances with modality missing individually, while the fully-paired multi-modal data is still represented by the unified hashing features. The cross-modal retrieval tasks including 'Image query Text' and 'Text query Image' are performed on Pascal VOC 2007 and NUS-WIDE. Assumed the database is fully-paired multi-modal instances while the query data is the unpaired ones. Figures 4 and 5 show the comparative experimental results on Pascal VOC 2007 and NUS-WIDE, respectively. As shown in Figs. 4 and 5, we can see that our method achieves comparable performance to the best baseline when the hash code length is set to 64 bit and 16 bit on Pascal VOC 2007

Table 3 mAP Comparison of multi-modal retrieval for different bits on NUS-WIDE

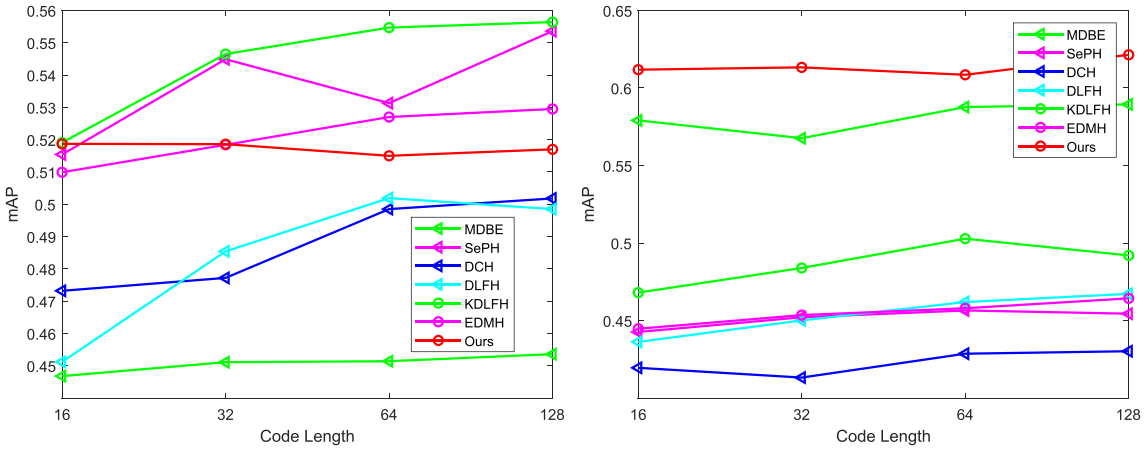
Methods	NUS-WIDE			
	16	32	64	128
ITQ [4]	0.3724	0.3751	0.3776	0.3789
LSH [3]	0.3421	0.3554	0.3544	0.3672
DLLE [6]	0.3738	0.3782	0.3794	0.3823
HCOH [11]	0.3232	0.3451	0.3434	0.3645
MFH [21]	0.3673	0.3752	0.3803	0.3815
MVLH [20]	0.3363	0.3339	0.3324	0.3284
FOMH [16]	0.3863	0.4450	0.5244	0.5519
OMH-DQ [17]	0.5223	0.5381	0.5823	0.5957
EPAMH [35]	0.4025	0.3969	0.3915	0.3879
FDCMH [34]	0.5632	0.5820	0.6018	0.6201
SIDMH [15]	0.5828	0.5976	0.6055	0.6120
Ours	**0.6190**	**0.6240**	**0.6271**	**0.6385**

Bold entries signify the best result

Table 4 mAP Comparison of multi-modal retrieval for different bits on IAPR-TC12

Methods	IAPR-TC12			
	16	32	64	128
ITQ [4]	0.3730	0.3844	0.3936	0.4020
LSH [3]	0.3251	0.3363	0.3509	0.3686
DLLE [6]	0.3644	0.3796	0.3863	0.3868
HCOH [11]	0.3082	0.3581	0.3717	0.3712
MFH [21]	0.3263	0.3374	0.3435	0.3451
MVLH [20]	0.3394	0.3401	0.3409	0.3499
FOMH [16]	0.3856	0.4083	0.4167	0.4291
OMH-DQ [17]	0.3949	0.4200	0.4446	0.4642
EPAMH [35]	0.3555	0.3874	0.3965	0.4088
FDCMH [34]	0.4012	0.4310	0.4385	0.4599
SIDMH [15]	04131	0.4277	0.4364	0.4706
Ours	**0.4198**	**0.4374**	**0.4571**	**0.4887**

Bold entries signify the best result

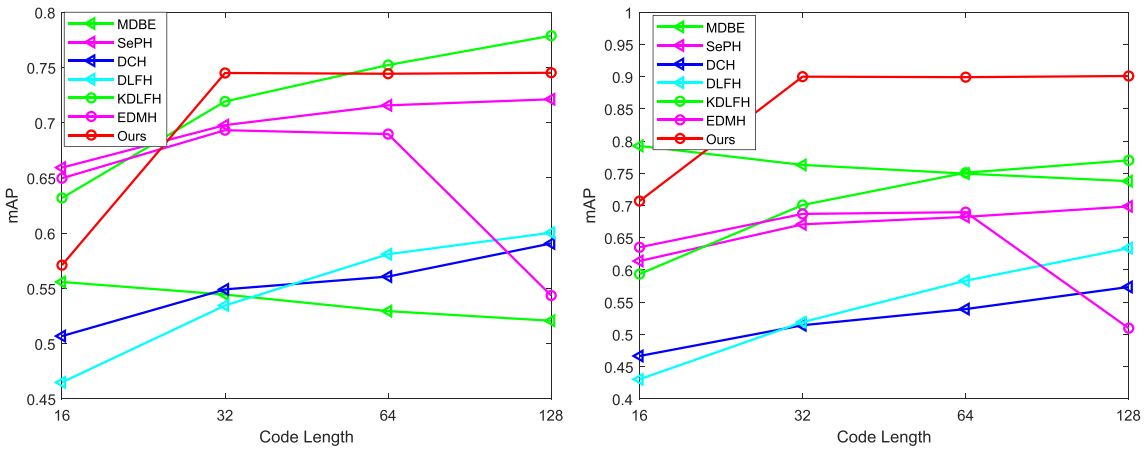


(a) Image query Text (b) Text query Image

Fig. 4 The comparison results for cross-modal retrieval on NUS-WIDE, (a) Image query Text, (b) Text query Image

and NUS-WIDE, respectively. On the Text query Image task, AMFH achieves much higher mAP scores than all comparison methods. The performance on the Text query Image task is superior to that on the Image query Text task. Furthermore, we set the proportion of paired instances in the database as 0.1, 0.3, 0.5, 0.7, and 0.9 respectively to observe the performance variation. The experimental results on Pascal VOC 2007 and NUS-WIDE are recorded in Table 5. It is easy to find that the variation performance on the Image query Text task is smaller than that on the Text query Image task with the increase of the proportion of paired samples in the database. The possible reason is that text modality is much better to represent the high-level semantic content of the described object than image modality.

6) *Ablation study:* In our framework, the projection matrices are learned during the training stage and the hash codes of newly coming multi-modal instances are generated in an online learning mode in order to capture the modality variations. We conduct some ablation experiments to validate the effectiveness of the proposed adaptive online strategy. Figure 6 shows the visually dynamic variation of each modality weight for new batch instances. The larger variation amplitude means that the difference among multi-modal data is larger. In Fig. 6, we can observe that the variation on WiKi dataset is smaller than that on other datasets. Let 'fixed' indicate the case where the weights learned in the training stage are applied to generate hash codes for newly coming data. As shown in Fig. 7, our method with dynamic weights exhibits a considerable improvement over 'fixed'. It is apparent that the improvement extent of our method on multi-modal retrieval precision is correlated with the variation amplitude of the dynamic weights. This implies that the self-weighting strategy is available to deal with the complex data scenarios. It impacts on the performance beneficially, especially for data with large diversity.



(a) Image query Text (b) Text query Image

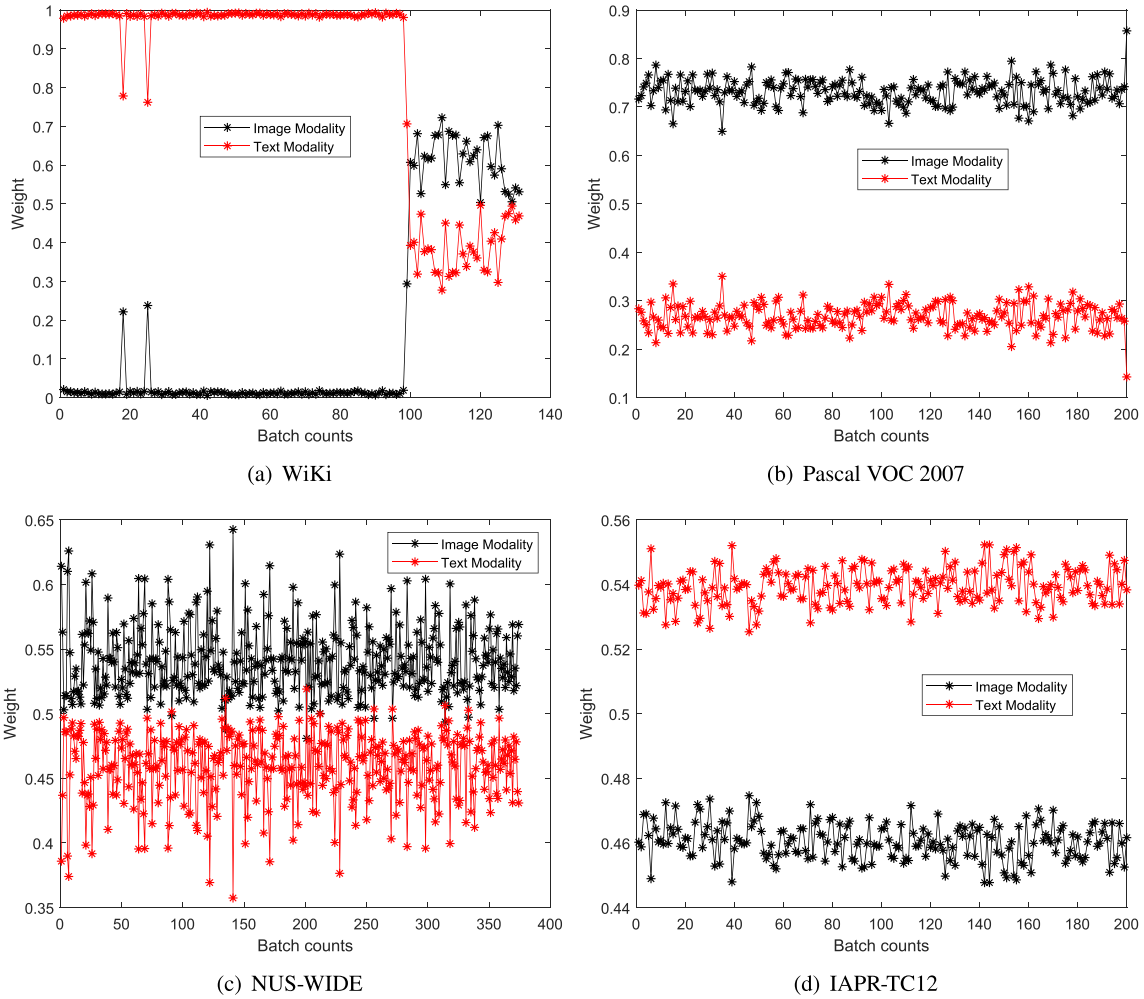
Fig. 5 The comparison results for cross-modal retrieval on Pascal VOC 2007, (a) Image query Text, (b) Text query Image

Table 5 The cross-modal retrieval results under the different proportion setting of paired samples

Dataset	Tasks	The proportion of paired samples				
		0.1	0.3	0.5	0.7	0.9
Pascal VOC 2007	Image query Text	0.7437	0.7442	0.7446	0.7449	0.7450
	Text query Image	0.7391	0.7791	0.8181	0.8516	0.8858
NUS-WIDE	Image query Text	0.5029	0.5058	0.5079	0.5103	0.5125
	Text query Image	0.4726	0.5139	0.5450	0.5770	0.6059

4.4 Complexity analysis

In this subsection, the complexity analysis of our proposed AMFH is provided. The computational complexity of the non-linear feature calculation for M modalities is $\mathcal{O}(Mnp)$. Updating $W^{(m)}$ and $\mu^{(m)}$ requires $\mathcal{O}(prn)$. The overall complexity during the projection learning stage is $\mathcal{O}(iter \times prn)$, where $iter$ denotes the number of iterations. Furthermore, we investigate the training time of our method and compare it with baselines by conducting experiments on WiKi, Pascal VOC 2007 and NUS-WIDE datasets respectively. The statistic results are reported in Table 6. As a popular data-independent method, the computation cost of LSH is relatively low. HCOH is a supervised method based on Hadam-ard matrix and its optimization process does not involve the matrix inverse operation. Except for LSH and HCOH, our method has the fastest training speed than the comparison methods. Although the training time of



(a) WiKi (b) Pascal VOC 2007

(c) NUS-WIDE (d) IAPR-TC12

Fig. 6 Visualization of dynamic modality weights on WiKi, Pascal VOC 2007, NUS-WIDE and IAPR-TC12

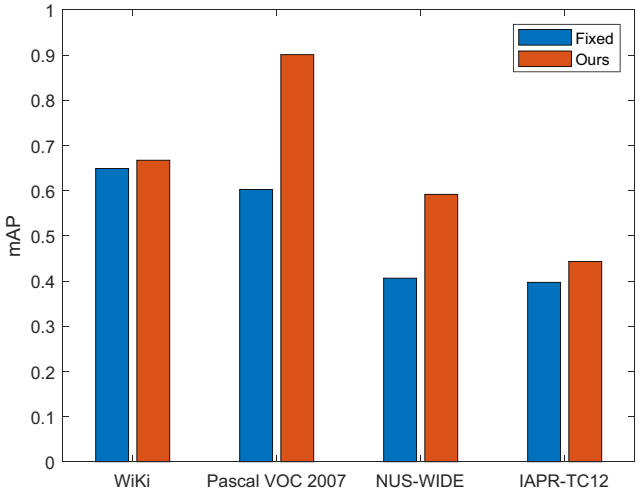


Fig. 7 The comparison results of the ablation experiments for the online hash encoding stage

our method is slightly higher than that of LSH and HCOH, it has much better performance compared with them. There is not the optimization process to discrete variables in AMFH, which improves the model training efficiency. It is obvious that our method requires the least training time than OMH-DQ, FOMH, FDCMH and SIDMH. The experimental results show that AM-FH greatly reduces the computation cost with the considerable retrieval performance.

4.5 Convergence analysis

The optimisation process based on the updating rule (see (12) and (14)) is decreasing the objective function monotonically, and rapidly converges to the minimum. This is shown by the experimental results on WiKi, Pascal VOC 2007, NUS-WIDE and IAPR-TC12. The convergence curves of our model with the code length of 128-bit on the three datasets are plotted in Fig. 8. The convergence curves for the other code length are similar to that for 128-bit. As seen in Fig. 8, our model converges within 5 iterations on four datasets.

4.6 Parameter sensitivity analysis

The penalty parameter δ in the objective function (11) is introduced to prevent model overfitting. We vary its values in the range of $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$ to explore the effect of δ on the performance of our model. The performance variation curves on four datasets are plotted in Fig. 9. We can find that a degradation commences on WiKi from $1e^{-3}$. In contrast, the performance is relatively stable for a large range of values on Pascal VOC 2007, NUS-WIDE and IAPR-TC12, which may be because the overfitting is less likely to happen on larger datasets. Our model is insensitive to the parameter and can flexibly be applied, especially to larger-scale multimedia retrieval problems. The empirical experiments are also conducted to observe the performance variations under different setting with parameter p. We tune its value from $\{100, 500, 1000, 1500, 2000\}$ and the experimental results are presented in Fig. 10. We see that the performance of our method is improved with the increase of p. Since the numerical value p is positively correlated with algorithm complexity, the determination of p must require comprehensive consideration of the retrieval performance and the training efficiency.

Table 6 Comparison of training time (seconds)

Methods	Training time (s)			
	WiKi	Pascal VOC 2007	NUS-WIDE	IAPR-TC12
ITQ [4]	0.5033	87.7270	2.5625	2.6890
LSH [3]	0.0129	2.9982	0.1014	0.2227
DLLE [6]	139.1312	147.1619	1461.2236	2728.1834
HCOH [11]	0.2302	9.9715	3.1992	4.9658
MFH [21]	2.7651	25.9216	19.0934	18.4506
MVLH [20]	184.7249	452.2433	913.1369	733.5932
FOMH [16]	1.2867	41.4450	3.1061	3.9645
OMH-DQ [17]	8.3657	192.0446	70.1140	22.7428
EPAMH [35]	–	94.1752	1.9401	3.3300
FDCMH [34]	73.4721	99.6400	165.2437	202.3413
SIDMH [15]	43.15121	204.3702	133.4563	189.3754
Ours	0.3724	23.8603	1.6507	1.0704

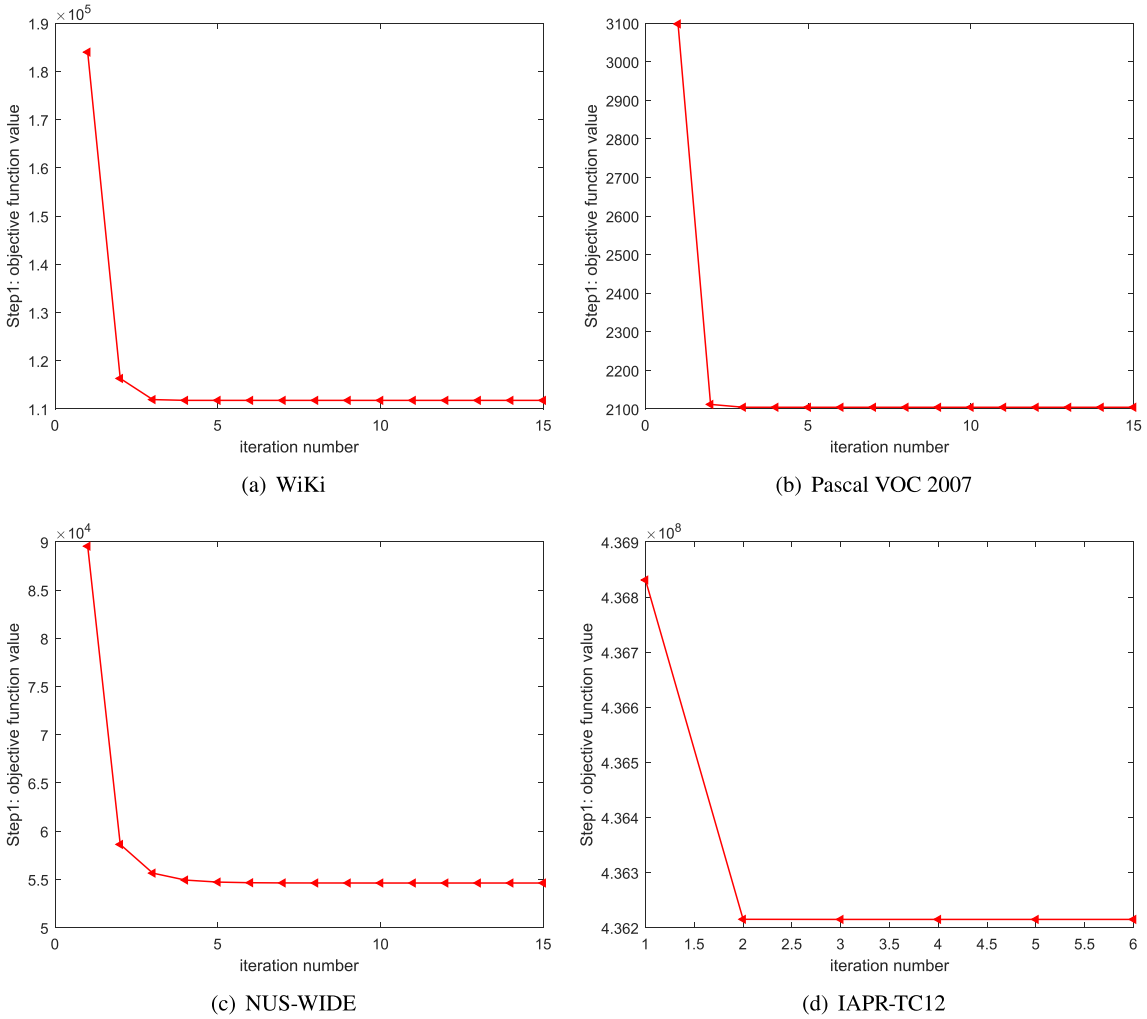


Fig. 8 Convergence curves on Wiki (a), Pascal VOC 2007 (b), NUS-WIDE (c) and IAPR-TC12 (d)

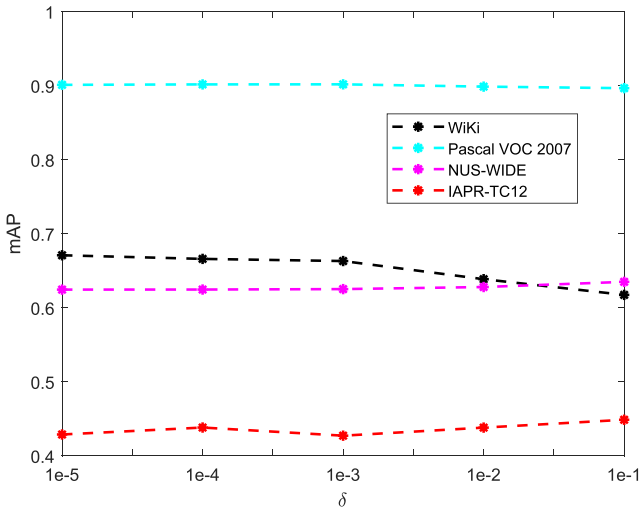


Fig. 9 mAP versus parameter δ

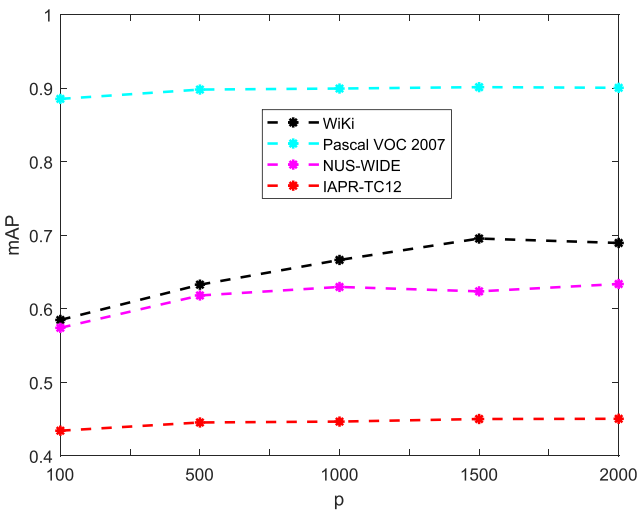


Fig. 10 mAP versus parameter p

5 Conclusion

In this paper, we proposed a novel multi-modal hashing method where Hadamard matrix is introduced to generate a discriminative hash center for each content category. Our model exhibits strong discriminative capability and is computationally light. As it is not highly sensitive to hyper-parameters, it can be applied flexibly. The extensive experiment results conducted on several public multi-modal datasets demonstrate the superior retrieval accuracy and efficiency of our proposed approach, as compared to the state-of-the-art methods.

The main limit of our proposed method is that the quantization error is larger compared with some supervised hashing methods. The overall optimisation problem is transform-ed into a relaxed problem without sign function. Although the relaxed problem is solved by an iterative optimization strategy, the information loss is inevitable. In the future, we plan to study a discrete optimization method to solve the problem with sign function.

Acknowledgements The authors would like to thank the anonymous reviewers for their encouragement and helpful comments. The paper is supported by the Research startup Fund project of Zhengzhou University of light industry (Grant No.2021BSJJ025), the Henan Provincial Department of Science and Technology Research Project (Grant No. 222102210064), and the National Natural Science Foundation of China (Grant No. 62162033).

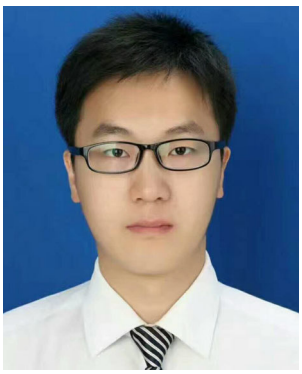
Declarations

Conflict of Interests 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.

References

1. Chen Y, Zhang H, Tian Z, Wang J, Zhang D, Li X (2020) Enhanced discrete multi-modal hashing: More constraints yet less time to learn. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2020.2995195
2. Chua TS, 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: Proceedings of the ACM international conference on image and video retrieval. https://doi.org/10.1145/1646396.1646452
3. Datar M, Immorlica N, Indyk P, Mirrokni VS (2004) Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth annual symposium on computational geometry, pp 253–262
4. Gong Y, Lazebnik S, Gordo A, Perronnin F (2012) Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans Pattern Anal Mach Intell 35:2916–2929
5. Hu M, Yang Y, Shen F, Xie N, Hong R, Shen HT (2019) Collective reconstructive embeddings for cross-modal hashing. IEEE Trans Image Process 28:2770–2784
6. Ji R, Liu H, Cao L, Liu D, Wu Y, Huang F (2017) Toward optimal manifold hashing via discrete locally linear embedding. IEEE Trans Image Process 26:5411–5420
7. Jiang QY, Li WJ (2019) Discrete latent factor model for cross-modal hashing. IEEE Trans Image Process 28:3490–3501
8. Koutaki G, Shirai K, Ambai M (2018) Hadamard coding for supervised discrete hashing. IEEE Trans Image Process 27:5378–5392
9. Li Z, Tang J, Mei T (2019) Deep collaborative embedding for social image understanding. IEEE Trans Pattern Anal Mach Intell 41:2070–2083
10. Lin M, Ji R, Liu H, Sun X, Chen S, Tian Q (2020) Hadamard matrix guided online hashing. Int J Comput Vis 128:2279–2306
11. Lin M, Ji R, Liu H, Wu Y (2018) Supervised online hashing via hadamard codebook learning. In: Proceedings of the 26th ACM international conference on multimedia, pp 1635–1643
12. Lin Z, Ding G, Han J, Wang J (2016) Cross-view retrieval via probability-based semantics-preserving hashing. IEEE Trans Cybern 47:4342–4355
13. Liu H, Ji R, Wu Y, Hua G (2016) Supervised matrix factorization for cross-modality hashing. In: International joint conference on artificial intelligence, pp 1767–1773
14. Liu X, He J, Liu D, Lang B (2012) Compact kernel hashing with multiple features. In: Proceedings of the 20th ACM international conference on multimedia, pp 881–884
15. Lu X, Liu L, Nie L, Chang X, Zhang H (2020) Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval. IEEE Transactions on Multimedia
16. Lu X, Zhu L, Cheng Z, Li J, Nie X, Zhang H (2019) Flexible online multi-modal hashing for large-scale multimedia retrieval. In: Proceedings of the 27th ACM international conference on multimedia, pp 1129–1137
17. Lu X, Zhu L, Cheng Z, Nie L, Zhang H (2019) Online multi-modal hashing with dynamic query-adaption. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 715–724
18. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on multimedia, pp 251–260
19. Shen F, Shen C, Liu W, Shen HT (2015) Supervised discrete hashing. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), pp 37–45
20. Shen X, Shen F, Sun QS, Yuan YH (2015) Multi-view latent hashing for efficient multimedia search. In: Proceedings of the 23rd ACM international conference on multimedia, pp 831–834
21. Song J, Yang Y, Huang Z, Shen H (2013) Effective multiple feature hashing for large-scale near-duplicate video retrieval. IEEE Trans Multimed 15:1997–2008
22. Sylvester J (1867) Lx. thoughts on inverse orthogonal matrices, simultaneous signsuccessions, and tessellated pavements in two or more colours, with applications to newton's rule, ornamental tile-work, and the theory of numbers. Lon Edinb Dublin Philos Mag J Sci 34:461–475. https://doi.org/10.1080/14786446708639914
23. Wang D, Gao X, Wang X, He L, Yuan B (2016) Multimodal discriminative binary embedding for large-scale cross-modal retrieval. IEEE Trans Image Process 25:4540–4554
24. Wang D, Wang Q, Gao X (2018) Robust and flexible discrete hashing for cross-modal similarity search. IEEE Trans Circuits Syst Video Technol 28:2703–2715
25. Wang J, Shen HT, Song J, Ji J (2014) Hashing for similarity search: A survey. arXiv: Data Structures and Algorithms
26. Wei Y, Zhao Y, Lu C, Wei S, Liu L, Zhu Z, Yan S (2017) Cross-modal retrieval with cnn visual features: A new baseline. IEEE Trans Syst Man Cybern 47:449–460

27. Xiaobo S, Fumin S, Li L, Yun-Hao Y, Weiwei L, Quan-Sen S (2018) Multiview discrete hashing for scalable multimedia search. ACM Trans Intell Syst Technol (TIST) 9:1–21
28. Xu X, Shen F, Yang Y, Shen HT, Li X (2017) Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans Image Process 26:2494–2507
29. Yi Z, Yeung DY (2012) Co-regularized hashing for multimodal data. In: International conference on neural information processing systems
30. Yu J, Wu X, Kittler J (2019) Discriminative supervised hashing for cross-modal similarity search. Image Vis Comput 89:50–56
31. Yuan L, Wang T, Zhang X, Tay FE, Jie Z, Liu W, Feng J (2020) Central similarity quantization for efficient image and video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3083–3092
32. Zhang D, Li WJ (2014) Large-scale supervised multimodal hashing with semantic correlation maximization. In: Proceedings of the Twenty-Eighth AAAI conference on artificial intelligence, AAAI Press, pp 2177–2183
33. Zhang D, Wu XJ, Yu J (2021) Learning latent hash codes with discriminative structure preserving for cross-modal retrieval. Pattern Anal Applic 24:283–297
34. Zheng C, Zhu L, Lu X, Li J, Cheng Z, Zhang H (2019) Fast discrete collaborative multi-modal hashing for large-scale multimedia retrieval. IEEE Trans Knowl Data Eng 32:2171–2184
35. Zheng C, Zhu L, Zhang S, Zhang H (2020) Efficient parameter-free adaptive multi-modal hashing. IEEE Signal Process Lett 27:1270–1274
36. Zhu L, Lu X, Cheng Z, Li J, Zhang H (2020) Flexible multi-modal hashing for scalable multimedia retrieval. ACM Trans Intell Syst Technol (TIST) 11:1–20



Jun Yu received his Ph.D. degree in Pattern Recognition from Jiangnan University and joined the College of Computer and Communication Engineering, Zhengzhou University of Light Industry in 2021. His research interests include multimedia information retrieval, computer vision and deep learning.

Donglin Zhang is currently pursuing toward the Ph.D. degree at the school of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China. His current research interests include multimedia information retrieval and pattern recognition.

Zhenqiu Shu received his Ph.D. degree in computer applications from Nanjing University of Science and Technology in 2015. He is currently an associate professor in the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His research interests include image processing, computer vision, and machine learning.

Feng Chen is currently a lecturer with the School of Computer Science and Technology, Anhui University of Technology, China. His research interests include computer vision, machine learning, and pattern recognition.