Hindawi
Security and Communication Networks
Volume 2022, Article ID 9638683, 12 pages
https://doi.org/10.1155/2022/9638683

Research Article

Cross-Modal Discrimination Hashing Retrieval Using Variable Length

Chao He,[1] **Dalin Wang**,[2] **Zefu Tan**,[1] **Liming Xu**,[3] **and Nina Dai**[1]

[1]*School of Electronic and Information Engineering, Chongqing Three Gorges University, Chongqing 404100, China*
[2]*Chongqing Preschool Education College, Chongqing 404047, China*
[3]*School of Computer Science, China West Normal University, Nanchong 637002, Sichuan, China*

Correspondence should be addressed to Nina Dai; dainina83@163.com

Received 20 April 2022; Accepted 10 August 2022; Published 9 September 2022

Academic Editor: Qing Yang

Fast cross-modal retrieval technology based on hash coding has become a hot topic for the rich multimodal data (text, image, audio, etc.), especially security and privacy challenges in the Internet of Things and mobile edge computing. However, most methods based on hash coding are only mapped to the common hash coding space, and it relaxes the two value constraints of hash coding. Therefore, the learning of the multimodal hash coding may not be sufficient and effective to express the original multimodal data and cause the hash encoding category to be less discriminatory. For the sake of solving these problems, this paper proposes a method of mapping each modal data to the optimal length of hash coding space, respectively, and then the hash encoding of each modal data is solved by the discrete cross-modal hash algorithm of two value constraints. Finally, the similarity of multimodal data is compared in the potential space. The experimental results of the cross-model retrieval based on variable hash coding are better than that of the relative comparison methods in the WIKI data set, NUS-WIDE data set, as well as MIRFlickr data set, and the method we proposed is proved to be feasible and effective.

1. Introduction

With the advent of the big data era, the different types of modal data, e.g., text, image, and audio for the Internet of Things and Mobile Edge Computing, are dramatically increasing [1]. The traditional single-mode data retrieval methods, e.g., text retrieval text, image retrieval image, and audio retrieval audio, are gradual shift to cross-modal retrieval, e.g., text retrieval image, text retrieval audio, image retrieval text, which makes the retrieval return with the characteristics of diverse information and rich content [2]. Over the last few years, the cross-modal retrieval algorithms have been recently receiving significant attention and progress due to the application research of guaranteed data privacy and privacy-preserving cooperative object classification [3, 4].

There are two main categories in these research methods. One is the potential subspace learning-based method [5–8], among which the canonical correlation analysis (CCA) is the most commonly used model [5]. The CCA mapped the two-modal data into a potential subspace to achieve the correlation maximization of the associated data pairs, and then directly retrieves the similarity query in the subspace. Given the paramount idea of the correlation maximization of relevant data in subspace, some experts have proposed other deformation model algorithms similar to the CCA model. Fu et al. proposed the generalized Multiview analysis (GMA) to maximize the subspace correlation of multimodal data and achieve the class-discriminant via adding label information, which is conducive to further boosting the accuracy of the cross-modal retrieval [6]. Costa Pereira et al. first projected the original feature data of each mode into their respective semantic feature space, and then mapped the semantic features of multimodes into a unified subspace via applying CCA or kernel CCA. The proposed model utilized the label information of the data to improve the classification area analysis, meanwhile avoiding the direct mapping of the original multimodal features into the unified subspace so

that the cross-modal retrieval performance is notably improved [7]. Mandal and Biswas proposed the generalized dictionary pair algorithm and achieved good results via learning unified sparse coding subspace [8]. Although some progress has been made in unified subspace learning-based cross-modal retrieval algorithms, there are still some problems in cross-modal retrieval of large-scale multimodal data scenarios, e.g., high computational cost, high data storage resource consumption, and weak stationarity. Therefore, another kind of cross-modal retrieval algorithm based on hashing coding has stimulated a lot of interest in the research community.

With the characteristics of storage consumption and efficient retrieval speed, the Hash coding technology is very suitable for large-scale data trans-modal and trans-media tasks, e.g., real-time multimodal data personalized recommendation [9], hot topic detection, and trans-media retrieval. In the Hash coding-based cross-modal retrieval method [10–13], for maintaining the connection between multimodal data, the multimodal data was projected into low-dimensional Hamming space through linear mapping, and then an XOR operation was performed to measure the similarity distance. Thus, the speed problem of large-scale data retrieval was solved effectively. However, most of the prior arts are only suitable for scenarios of the single label and paired training data. Therefore, Mandal et al. first proposed a hashing cross-modal retrieval model for multiple training scenarios [14]. However, this model is similar to the method presented in Refs. [15, 16] that maps multimode data into equal-length hash coding, so that the data of various modes may not be well represented. In addition, the solution of binary hash coding is an NP-hard problem, which relaxes the binary constraint of hash coding, so that the learned hash coding is not accurate enough. For analytical simplicity, this paper first proposed a cross-modal retrieval model based on variable-length hash coding and added binary constraints in the process of solving hash coding. Therefore, the learned variable-length hash coding can better represent the original multimodal data and achieve higher accuracy. The main highlights of this paper are organized as follows.

(1) To combat the issue caused by the same length, we propose a variable-length hash coding-based cross-modal retrieval model in this paper, i.e., all modal data are projected into the hash coding space of the optimal lengths. Therefore, compared with the hash coding space of the fixed length, the original multimodal data can be represented more easily, and the model in this paper is more flexible in debugging experiments.

(2) We propose a more generalized multiscene cross-modal retrieval. The great majority of the existing cross-modal retrieval models, based on single label and pairwise multimodal dataset scenarios, cannot be applied to multilabel and unpaired multimodal dataset scenarios. In addition, the cross-model retrieval in this paper has good adaptability to single label or multilabel, paired, or unpaired multimodal dataset scenarios.

(3) Based on the single-modal data hash method, we propose a variable-length discrete hash coding-based cross-modal retrieval algorithm, and the validity of the algorithm is verified on several public data sets.

2. Related Works

This section mainly introduces several related hash coding cross-modal retrieval algorithms, which are also served as benchmark algorithms in the experimental process. Any reader who has a great interest in other cross-modal retrieval models, such as incorporating feedback technology and deep learning, can refer to Ref. [17].

2.1. Hashing Cross-Modal Retrieval Based on Semantic Correlation Maximization. Taherkhani et al. proposed a Semantic Correlation Maximization (SCM)-based cross-modal hash retrieval model. Meanwhile, compared with other supervised hash cross-modal retrieval models, this model has the advantages of lower training time complexity, better adaptability, and more stability for large-scale data sets [10]. The main highlights are as follows. (1) The calculation of the complex pin-to-pair similarity matrix can be avoided directly via applying label information of the training data set to calculate the similarity matrix, thus only small linear time complexity can be achieved, which also makes the model more stable. (2) The serialization solution method of hash coding is proposed via the computation code of bit by bit on the closed interval. Therefore, there is no need to set hyperparameters and stop conditions. To use label semantic information, cosine similarity between label vectors is used to construct the similarity matrix, and the similarity between the data object i and the data object j is defined as follows.

$$S_{ij} = \frac{\langle l_i, l_j \rangle}{\|l_i\|_2 \|l_j\|_2}, \tag{1}$$

where $\langle l_i, l_j \rangle$ represents the inner product of the corresponding label vector and $\|l\|_2$ describes the binary norm of the label vector. To achieve a cross-modal similarity query, the hash function should maintain the semantic similarity of multimodal data. More specifically, the hash coding of each modal data can reconstruct the semantic similarity matrix. The specific objective function of the SCM model is defined as follows:

$$\min_{W_x, W_y} \left\| \operatorname{sign}(XW_x) \operatorname{sign}(YW_y) - cS \right\|_F^2, \tag{2}$$

where X and Y represent the data of the two modes, W defines the linear transformation matrix, c describes the equilibrium parameter, and S defines the similarity measurement between two data among different modalities. There is a symbolic function in (2), so it is obvious that the optimization solution is an NP-hard problem, which relaxes the constraints of the symbolic function and adds the constraints between the bits of the hash coding. Finally, the transformation matrixes W_x, W_y of each modal data can be calculated, so that the hash coding of new data can be resolved.

2.2. Hashing Cross-Modal Retrieval Based on Semantic Preserving. Chen et al. proposed a Semantic Preserving Hash cross-modal retrieval (SEPH) model, which converts the similar association information of data into the form of the probability distribution and then approximates hash coding via minimizing the Kullback–Leibler (KL) divergence distance [11]. The whole objective function model is effectively guaranteed in mathematical theory. As with the SCM model, the similarity matrix is first constructed to provide supervisory information for the learned hash coding. This model mainly includes two steps, i.e., hash coding solution and learning of kernel logic Sti regression function. When it comes to the process of solving the hash coding, the similarity matrix is first transformed into the form of probability P, and the semantic probability distribution Q on the unified hash coding is calculated, then the KL distance between the two distributions is minimized, and the semantic preserving hash coding is resolved.

$$P_{ij} = \frac{S_{ij}}{\sum_{i \neq j} S_{ij}},$$
$$Q_{ij} = \frac{\left(1 + h\left(B_i, B_j\right)\right)^{-1}}{\sum_{t=1} \left(1 + h\left(B_i, B_t\right)\right)^{-1}}, \tag{3}$$

where $h(,)$ represents the Hamming distance function of hash coding; learning the best hash coding B aims to make the distribution between P and Q as similar as possible. The KL distance between the distributions is measured as follows:

$$D_{KL}(PQ) = \sum_{i \neq j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right). \tag{4}$$

In all, a better unified semantically preserving hash coding can be calculated according to the solution steps, and then the logistic regression mapping function of each modal data mapped to the unified hash coding is learned. The representation of learning the $k(1 \leq k \leq K)$-th Logistic regression function for X mode data is defined as follows:

$$\min_{w^k} \sum_{i=1}^{n} \log\left(1 + e^{-b_i^k x_i w^k}\right) + \lambda \left\| w^k \right\|_2^2, \tag{5}$$

where $b_i^k \in \{-1, +1\}^{n \times 1}$ defines the column vector on the k-th bit attribute of the common binary code, and the transformation matrix w^k can be solved. Then, the probability that the value b belongs to -1 and $+1$ at the k-th bit of the binary code of the new sample x^q data in X mode can be calculated as follows:

$$P\left(c = b | x^q\right) = \left(1 + e^{-b x^q w^k}\right)^{-1}. \tag{6}$$

Therefore, the value at the k-th bit of data binary coding is selected as the value corresponding to the high probability, which is defined as follows:

$$c^k = \operatorname{sign}\left(P\left(c = 1 | x^q\right) - P\left(c = -1 | x^q\right)\right). \tag{7}$$

Finally, the k-th logistic regression function on the X mode data can be learned, and then the new sample x^q is mapped into the binary coding with the growing degree of K. The final hash coding can be achieved by changing the element with the value of -1 into 0.

2.3. Hashing Cross-Modal Retrieval Based on Generalized Semantic Preserving. Because most of the existing cross-modal retrieval methods require multimodal data to appear in pairs, i.e., another modal data corresponding to text or image exists in training set data, Mandal et al. proposed a Generalized Semantic Preserving Hashing model (GSPH) for N-label cross-modal retrieval, which is suitable for a single label or multilabel, paired or unpaired multimodal data application scenarios [14]. The GSPH model first learns the optimal hash coding of each modal data, meanwhile the hash coding preserves the semantic similarity between the multimodal data and then learns the hash function of multimodal data mapped to the hash coding space. The main highlights are as follows. (1) A hash model that can deal with single-label paired data and single-label unpaired data is proposed for the first time. (2) The generalized hash cross-modal retrieval model is proposed, which can be applied to the scenarios of single-label paired data, single-label unpaired data, multilabel paired data, as well as single-label unpaired data. Meanwhile, the semantic similarity of data is maintained by the common hash coding. As with SCM and SEPH methods, the GSPH algorithm also needs to define the similarity matrix $S \in R^{N_1 \times N_2}$ between multimodal data, where N_1 and N_2 are the sample numbers of X and Y modal data, respectively, so the objective function of the GSPH model is defined as follows:

$$\min \left\| S - \left(\frac{1}{q}\right) B_x B_y^T \right\|_F^2 s.t B_x \in \{-1, +1\}^{N_1 \times q}, B_y \in \{-1, +1\}^{N_2 \times q}. \tag{8}$$

The binary coding B_x and B_y of the X and Y modal data can be calculated by the GSPH method, and then the mapping function of the original data for each modal into hash coding needs to be learned. Just like the SEPH method, the logistic regression function is selected as the mapping function. Therefore, readers can refer to Section 2.2 for learning the mapping hash function and generating the hash coding of new samples.

3. Cross-Modal Retrieval Based on Variable-Length Hash Coding

In this section, the cross-modal retrieval algorithm of variable-length hash coding is presented, and the optimization process of the objective function and time complexity of the algorithm is analyzed. To facilitate the analytical simplicity and reduce the experimental operation, this paper mainly studies the case of two-modal data and gives the algorithm model extended to three or more modal data in Section 3.5.

3.1. Algorithmic Model. The variables presented in this paper are defined as follows. $X \in R^{d_1 \times n_1}$ and $X \in R^{d_2 \times n_2}$ represent the original feature data sets of the two modes, respectively, $B_X \in R^{q_1 \times n_1}$ and $B_Y \in R^{q_2 \times n_2}$ are the corresponding variable-

length hash coding, where each column represents a sample and each row represents attribute features. In addition, P_X and P_Y are the projection matrixes, and W is the association matrix of two modes. The similarity matrix $S \in R^{n_1 \times n_2}$ between multimode data is constructed as follows:

$$S_{ij} = \begin{cases} \langle l_x^i, l_y^j \rangle & I, \\ e^{-\left\| l_x^i - l_y^j \right\|^2 / \delta} & II, \\ 1, \text{if} l_x^i = l_y^j; 0, if l_x^i \neq l_y^j & III, \end{cases} \tag{9}$$

where l defines the label vector of the sample, and each element S_{ij} of the similarity matrix represents the similarity between X modal data i and Y modal data j. The next goal of this paper is to learn the compact hash coding of the optimal length for each model, so that these hash coding can perfectly represent the original multimode data and maintain the semantic similarity of multimode data sets. This paper calculates the similarity of different modal data in potential space by referring to Ref. [7] and assumes that there is a common potential abstract semantic space V between multimodal data, in which multimodal data can be queried and retrieved directly. And, each modal hash coding is projected into the potential abstract semantic space in the following form:

$$M_1 \colon B_X \xrightarrow{W_1} V_X M_2 \colon B_Y \xrightarrow{W_2} V_Y. \tag{10}$$

In the space V, the similarity between data can be calculated according to the relation of the inner product, which is defined as follows:

$$\widetilde{S} = V_X^T V_Y = \left(W_1 B_X\right)^T \left(W_2 B_Y\right) = B_X^T W_1^T W_2 B_Y. \tag{11}$$

Remembering $W = W_1^T W_2$, we do not need to explicitly solve the existing form of each mode data in the potential abstract semantic space V, but only calculate the similarity W between the varied-length hash coding of each mode. The cross-modal retrieval objective function of the specific variable-length hash coding is defined as follows:

$$\begin{aligned} &\min_{B_X, B_Y, W, P_X, P_Y} \left\| B_X - P_X X \right\|_F^2 + \left\| B_Y - P_Y Y \right\|_F^2 \\ &+ \left\| S - B_X^T W B_Y \right\|_F^2 s.t. B_X \in [-1, +1]^{q_1 \times n_1}, B_Y \in [-1, +1]^{q_2 \times n_2}. \end{aligned} \tag{12}$$

The first two terms of (12) are applied to, respectively, project the two-modal data into the hash coding space of the optimal lengths, and the last term indicates that the variable-length hash coding in the potential space still maintains the semantic similarity relation of the original multimodal data. The corresponding projection matrixes P_X, P_Y, hash coding B_X, B_Y, and correlation matrix W can be solved simultaneously through optimization.

3.2. Model Solution Procedure. To simplify the difficulty of solving hash coding, the prior art converts binary constraint conditions of hash coding into solving continuous real-valued problems and then obtains approximate hash coding through symbolic functions [10–12]. However, the solved hash coding has essential defects and cannot represent the original multimodal data effectively. The binary constraint condition of hash coding is always maintained in the solving process of this subsection. When the objective function is solved, the variables B_X, B_Y, W, P_X, P_Y of simultaneous solution are nonconvex and difficult to solve. Therefore, this paper first solves one of the variables and fixed the remaining variables, and then solves the other variables in this way. All variables are solved by iteration until the objective function tends to converge.

(a) Fix other variables and resolve P_X, P_Y. Therefore, the objective function can be simplified in the following form:

$$\min_{P_X} \left\| B_X - P_X X \right\|_F^2 \min_{P_Y} \left\| B_Y - P_Y Y \right\|_F^2. \tag{13}$$

Therefore, the analytical formulae can be calculated by regression formula, respectively,

$$\begin{aligned} P_X &= B_X X^T \left(X X^T\right)^{-1}. \\ P_Y &= B_Y Y^T \left(Y Y^T\right)^{-1}. \end{aligned} \tag{14}$$

(b) Fix other variables and resolve W. The objective function can be simplified in the following form:

$$\min_W \left\| S - B_X^T W B_Y \right\|_F^2. \tag{15}$$

It is obvious that (15) is a bilinear regression model, and the analytical formula is as follows:

$$W = \left(B_X B_X^T\right)^{-1} B_X S B_Y^T \left(B_Y B_Y^T\right)^{-1}. \tag{16}$$

(c) Fix other variables and resolve B_X. The objective function can be simplified in the following form:

$$\begin{aligned} &\min_{B_X} \left\| B_X - P_X X \right\|_F^2 + \left\| S - B_X^T W B_Y \right\|_F^2 s.t. B_X \\ &\in [-1, +1]^{q_1 \times n_1}. \end{aligned} \tag{17}$$

Because of the two-value constraint, it is complicated to resolve directly. Therefore, in this paper, the variable B_X is solved successively, i.e., when solving a row vector of B_X, the remaining row vectors are fixed first, and then the other row vectors are solved iteratively. (17) can be further transformed into (18).

$$\begin{aligned} &\min_{B_X} \left\| B_X \right\|_F^2 - 2Tr\left(B_X^T P_X X\right) + \left\| P_X X \right\|_F^2 + \left\| S \right\|_F^2 \\ &- 2Tr\left(B_X^T W B_Y S^T\right) + \left\| B_X^T W B_Y \right\|_F^2 s.t. B_X \\ &\in [-1, +1]^{q_1 \times n_1}. \end{aligned} \tag{18}$$

Because of the binary constraint, it is obvious that the first term is a constant, i.e., $\left\| B_X \right\|_F^2 = q_1 * n_1$. If constant terms and irrelevant variables B_X are removed, (18) can be rewritten into a more concise form.

$$\min_{B_X} \left\| D B_X \right\|_F^2 - 2Tr\left(B_X^T Q\right) s.t. B_X \in [-1, +1]^{q_1 \times n_1}, \tag{19}$$

where $D = B_Y^T W^T$, $Q = (W B_Y S^T + P_X X)$ and $Tr(\ldots)$ are the trace of the solution matrix. After

deformation, the solution of (19) has a relationship with the solution of the objective function in Ref. [16], so this paper refers to its solution process. When solving the i-th row vector z^T of B_X, let B_X' be the matrix B_X after row vector deletion z^T, p^T defines the i-th row vector of Q, Q' represents the matrix Q after row vector deletion p^T, d defines the i-th column vector of D, and D' represents the matrix D after column vector deletion d, and then refer to the solution results in Ref. [16].

$$z = \operatorname{sign}\left(p - B_X' D'^T d\right). \tag{20}$$

The i-th row vector of B_X can be resolved, and then the remaining row vectors can be solved via a similar procedure.

(d) Fix other variables and resolve B_Y.

In the process of solving B_Y, it is similar to solving B_X, so readers can refer to the solution method of B_X for a detailed solution of B_Y.

3.3. Algorithm Description. To project hash coding into the optimal space for comparison, measurement, and retrieval, the associated transformation matrix W is introduced into the cross-modal retrieval model of variable-length hash coding on the base of the GSPH model, and then the similarity between data can be compared in the potential space through W. Subsection 2.2 provides the solution process of each variable in the model, and the overall training steps for the model are shown in Algorithm 1.

According to the proposed training process, the projection matrix of each mode can be calculated separately, and then the corresponding hash coding can be solved by a symbolic function. For query sample x' or y', the corresponding hash coding generation method is $b' = \operatorname{sign}(P_X x')$ or $b' = \operatorname{sign}(P_Y y')$. To improve the accuracy of generating corresponding hash coding, the query sample pair information (x', y') of these two modes can be used to generate hash coding simultaneously. If the final hash coding is expected to exist in the hash coding space of the X mode, then $b' = \operatorname{sign}(P_X x' + \theta W P_Y y')$. If the final hash code is desired to exist in the hash coding space of the Y mode, then $b' = \operatorname{sign}(P_Y y' + \theta W^T P_X x')$, where θ is a non-negative equilibrium parameter. The overall testing steps for the model are summarized in Algorithm 2.

3.4. Time Complexity. The time complexity of the cross-modal retrieval algorithm in this section is mainly composed of computation-related variables. In the training phase, the time of each iteration is consumed in updating the projection matrixes P_X, P_Y, transformation matrix W, and corresponding hash coding matrixes B_X, B_Y, in which these variables are calculated by (14) and (16), and (17), respectively, and the corresponding calculation time complexity is $O(d^2qn), O(q^2n^2), O(dq^2n)$. Therefore, the total time complexity of the proposed model is $O((d^2 + qn + dq)qnT)$, where T represents the total number of iterations, where $d = \max(d_1, d_2), q = \max(q_1, q_2), n = \max(n_1, n_2)$. More specially, d_1, q_1, and n_1 are the original dimension, hash length, and the total number of samples of X mode data, respectively, and d_2, q_2, and n_2 are the original dimension, hash length, and the total number of samples of Y mode data, respectively. Once the training process is end, the time and space complexity for generating a new sample is $O(dq)$.

3.5. Application Scenario. The cross-modal retrieval model can be easily extended to the scenarios of three or more modal data, assuming that $m\,(m > 2)$ modal data, then the cross-modal retrieval model of variable-length hash coding for m modal data is defined as follows:

$$\min_{B_i, W^{(i,j)}, P_i} \sum_{i=1}^{m} \left\| B_i - P_i X_i \right\|_F^2 + \sum_{i,j}^{m} \left\| S^{(i,j)} - B_i^T W^{(i,j)} B_j \right\|_F^2 \tag{21}$$
$$s.t. B_i \in [-1, +1]^{q_i \times n_i}.$$

The first item in (21) represents the hash code mapping of all modal data into the optimal length, and the second item represents the semantic relationship preservation between the hash coding of each mode and another modal hash coding. The process of model optimization and query sample hash coding generation can follow the way of two-modal data scenarios.

4. Results and Discussion

4.1. Data Sets and Performance Metrics. To verify the validity of the model, the commonly used WIKI data set, NUS-WIDE data set, and MIRFlickr data set are selected for the cross-modal retrieval. In addition, the precision-recall and Mean Average Precision (MAP) index are used to measure model performance as shown in Refs. [11–13].

WIKI data set is collated from Wikipedia page [7], and each image has the corresponding description text, in which each text contains no less than 70 words. The data sets belong to a single-label data, and there are 10 categories, each image or text belongs to one of these categories, and images or texts belonging to the same category are considered to have similar semantic information. There exist 2866 samples (2173 training sets and 693 test sets), in which image data is represented by 128-dimensional Scale Invariant Feature Transform (SIFT) features and text data by 10-dimensional Latent Dirichlet Allocation (LDA) features.

NUS-WIDE data set is collected and sorted from the Internet by the National University of Singapore [18], which regulates 269,648 images and explanatory annotations accomplished by about 5,000 people. Each sample belongs to multilabel data, which is eventually divided into 81 categories. Due to the sample numbers of some categories differ greatly in this paper, just as Refs. [10, 11], the top 10 categories with many samples are firstly selected, and finally 186,577 text-image pairs have been achieved. Text and image are considered similar, if there is at least one of the same category attributes. Subsequently, 1% of the data (about 1866) are randomly selected as the test set and 5000 samples as the training set. The images of the NUS-WIDE data set are

Input: Training datasets X/Y and label matrix L_X/L_Y; Initialized association matrix W; Initialized variable-length hash B_X, B_Y; Initialized iteration control parameter T
Output: Variables B_X, P_X, B_Y, P_Y, W
Procedure:
(0) Applying label matrix L_X, L_Y and (9) to construct a semantic similarity matrix S
(1) $iter = 0$;
(2) while $iter < T$ do
(3) According to (14), update the dictionary projection matrix P_X, P_Y;
(4) According to (16), update the association matrix W;
(5) According to equation (18) and the detailed solving process in Ref. [14], the hash code of variable length is updated one line at a time and finally updated as a whole B_X, B_Y;
(6) If the objective function (12) tends to converge, and stop the iteration; otherwise, skip to step (2);
(7) End while

Algorithm 1: Training produce of proposed method.

Input: Testing datasets X'/Y'; trained $f(\cdot)$, $g(\cdot)$ and $\mathbf{W}$.
Output: The top n cross-modal data matching the samples to be retrieved.
Procedure:
(1) if input independent x' or y' then
(2) compute the corresponding hash code by $b' = \text{sign}(f(x'))$ or $b' = \text{sign}(g(y'))$;
(3) end if
(4) if input paired $(x', y')x'$ then:
(5) if hash code exists in space of Y data:
$b' = \text{sign}(g(y')) + W^{\mathrm{T}} f(x')$;
(6) else:
$b' = \text{sign}(f(x')) + W^{\mathrm{T}} g(y')$;
(7) end if
(8) end if
(9) Calculate the Hamming distance between the hash code b' and the hash codes of all samples in the retrieval database
(10) Sort the distances calculated in ascending order, and return the first n samples.

Algorithm 2: Testing produce of the proposed method.

represented by 500-dimensional SIFT features and the text data by the word frequency of 1000 dimensions.

MIRFlickr data set originated from the Flickr website, which contains 25000 images and corresponding manually annotated text information [19]. Just as Ref. [11], we have deleted some data without labels or with less than 20 times of labeled words, and finally 16,738 samples are divided into 24 categories. Each image text pair belongs to multicategory data, which contains at least one category label. This paper selects 5% data as a test set and 5000 samples as the training set. Images in the data set are represented by 150-dimensional edge histograms and text by 500-dimensional vectors. The evaluation criteria are defined as follows:

$$\begin{aligned} \text{Accuracy: } P(N) &= \frac{n}{N} \times 100\%, \\ \text{Recall: } R(N) &= \frac{n}{N_r} \times 100\%, \end{aligned} \tag{22}$$

where n represents the number of relevant samples among N results stemming from the retrieval and N_r defines the number of samples related to query samples in the whole database.

Average Precision (AP) indicator calculation: Given a query sample and the first R returned results, the AP calculation equation of this sample is defined as follows:

$$AP = \frac{1}{K} \sum_{r=1}^{R} P(r)\delta(r), \tag{23}$$

where K represents the number of retrievably returned results related to query samples, and $P(r)$ defines the accuracy of the returned first r retrieval results. If the r-th retrieval result is related to the query sample, $\delta(r)$ is 1; otherwise, $\delta(r)$ is 0. Finally, the AP average value of all query samples is solved, which is the MAP index to evaluate the overall search performance.

4.2. Benchmark Algorithm. In this subsection, the various multimodal data are preprocessed according to the method represented in Ref. [16], i.e., the distance between sample points and randomly selected reference points is calculated. Then the discrete supervised hash model is used to initialize the hashing coding of each mode. To highlight the importance of the label matrix in the process of optimization, the

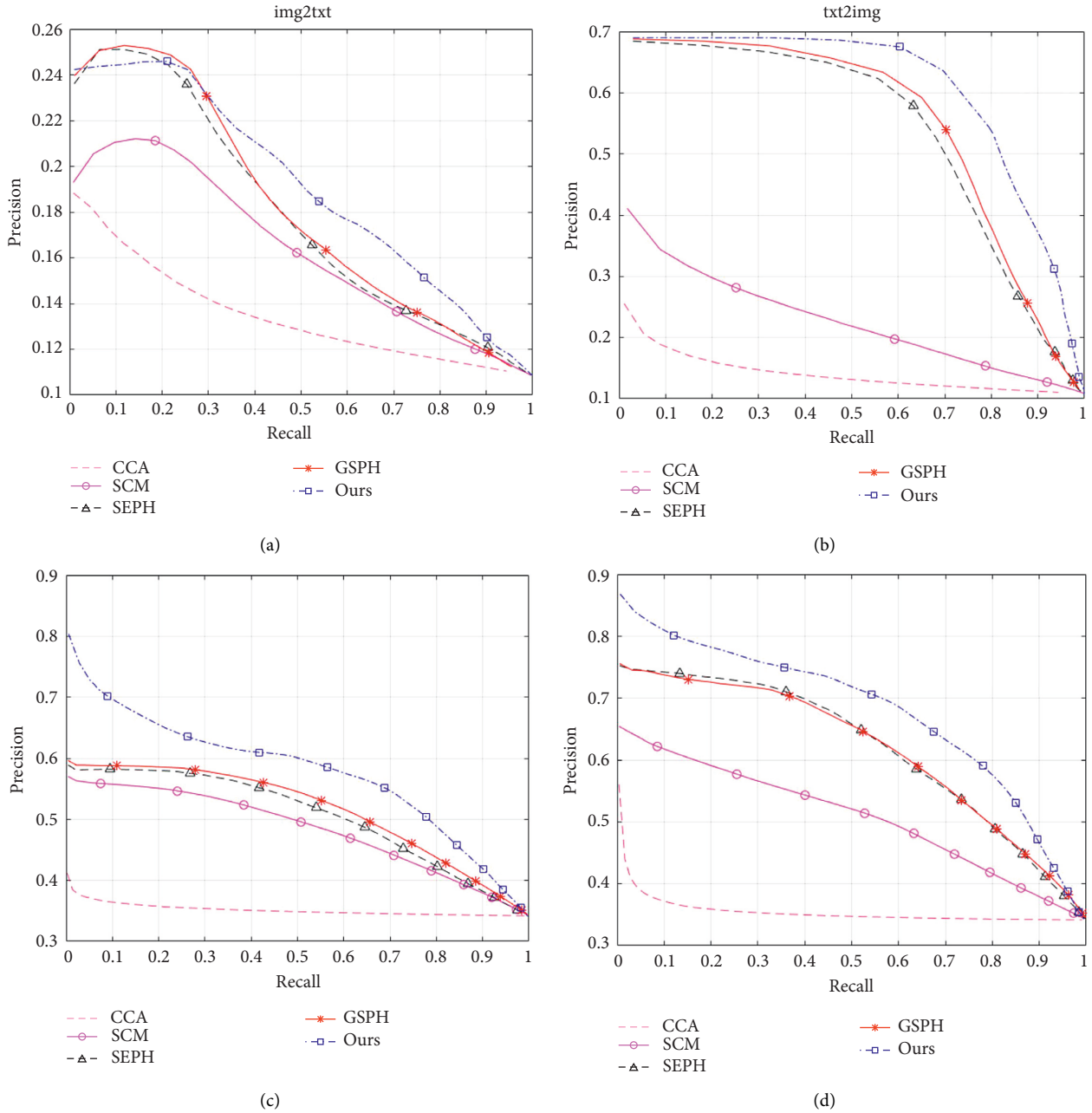


Figure 1: Continued.

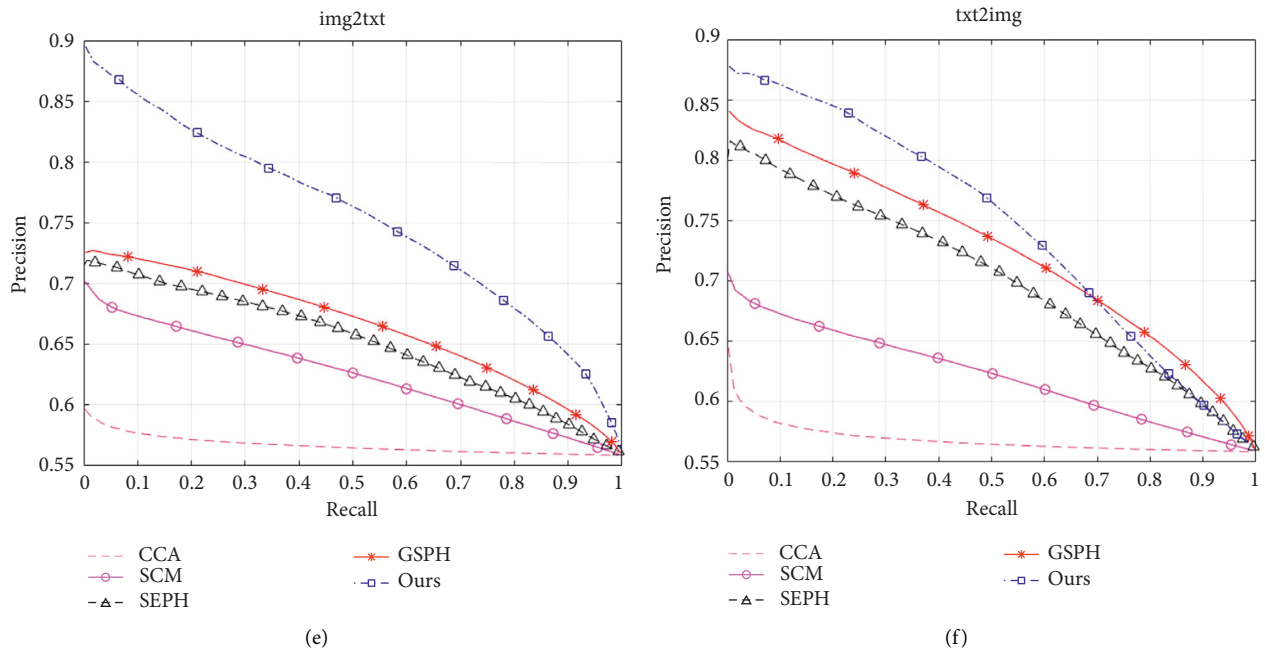


Figure 1: Precision rate and recall rate of different methods for the different data sets: (a) WIKI (img2txt), (b) WIKI (txt2img), (c) NUS-WIDE (img2txt), (d) NUS-WIDE (img2txt), (e) MIRFlickr (txt2img), and (f) MIRFlickr (txt2img).

label matrix of all data is enlarged by 10 times. In addition, CCA, a typical correlation analysis method commonly used in the field of cross-modal retrieval, and the cross-modal retrieval algorithm based on semantic correlation hash coding in recent years are selected as a comparative experiment. These hashing cross-modal retrieval models are SCM, SEPH, and GSPH, respectively, and the comparison experiments proposed in this paper are implemented in MATLAB with the help of the parameters set in the original text. Both SEPH and GSPH models include two methods to learn hash functions: (1) training hash functions SEPH_rnd and GSPH_rnd based on randomly selected samples; (2) training hash functions SEPH_knn and GSPH_knn based on selecting samples through clustering. The experiment shows that the performance of the hash function obtained by these two training methods is the same. Therefore, the first method, randomly selected samples, is selected to train the hash functions of both SEPH and GSPH models in the comparative experiment. Moreover, the two different methods in the SCM model are SCM_seq and SCM_orth, and the experiment results show that the former is generally superior to the latter; therefore the former is used as a comparative experiment [10].

4.3. Experimental Results. This subsection presents the experimental results of cross-modal retrieval on the WIKI dataset, NUS-WIDE dataset, and MIRFlickr dataset. The following cross-modal retrieval tasks include image retrieval text and text retrieval image, and these two retrieval tasks are analyzed in detail. Figure 1 shows the curves of retrieval accuracy rate and recall rate on three kinds of data sets. To facilitate the comparison with the benchmark algorithm, both image and text are projected into equal-length hash coding space (64 bits). It can be seen from Figure 1 that the performance of the method proposed in this paper is generally superior to that of the comparison method, although the front part of the curve (subgraph (a) of Figure 1) in the image retrieval text task on the WIKI dataset is slightly lower than that of SEPH and GSPH methods. However, it can be seen from the subgraph (a) of Figure 2 that the effect of the optimal hash coding combination length in this paper is slightly higher than that of SEPH and GSPH methods. It can also be seen from Figure 1 that for the other two groups of multilabel data, the effect of this paper has been improved more than that of the comparison methods, due to the model in this paper being more suitable for multilabel data sets than the CCA, SCM, SEPH, and GSPH models.

The MAP index of image retrieval text and text retrieval image of each method is presented in detail in Tables 1 and 2, respectively, and the highest MAP value of each column is marked black. To compare the effects of CCA and other methods, this paper projected data into subspaces of different dimensions to observe the influence of CCA methods. Tables 1 and 2 show that the MAP value of the proposed method and other hash coding methods increases slightly as the length of hash coding increases. As can be seen from the numerical part marked black in the table, the MAP value of the proposed method is superior to that of the comparison method, no matter in the image retrieval text task or the text retrieval image task. Given that the hash coding length is 64 bits, this paper improves about 15%, 10%, and 13% in the image retrieval text task on WIKI, NUS-WIDE, and MIRFlickr data sets, and about 12%, 11%, and 5% in the text retrieval image task compared with the GSPH method.

Figure 2 shows the experimental results of different length combinations for the hash coding proposed in this

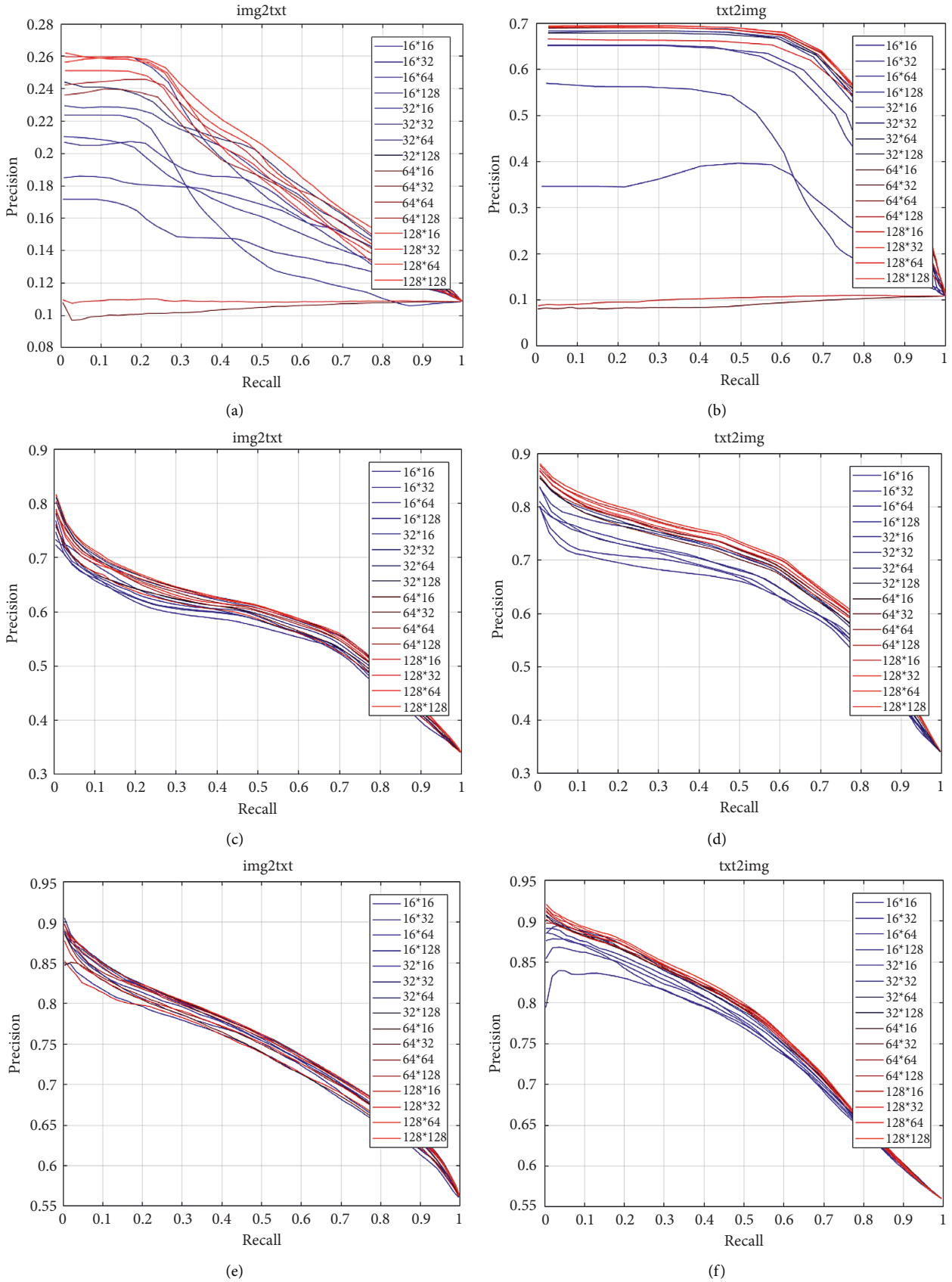


Figure 2: Precision rate and recall rate of different hash coding length combinations: (a) WIKI (img2txt), (b) WIKI (txt2img), (c) NUS-WIDE (img2txt), (d) NUS-WIDE (img2txt), (e) MIRFlickr (txt2img), and (f) MIRFlickr (txt2img).

Table 1: MAP image retrieval text img2txt.

	WIKI data set				NUS-WIDE data set				MIRFlickr data set			
	16	32	64	128	16	32	64	128	16	32	64	128
CCA	0.184	0.170	0.150	0.140	0.373	0.366	0.361	0.358	0.579	0.574	0.571	0.568
SCM	0.234	0.241	0.246	0.257	0.501	0.542	0.553	0.551	0.610	0.631	0.647	0.641
SEPH	0.276	0.296	0.300	0.313	0.560	0.578	0.582	0.581	0.671	0.652	0.681	0.648
GSPH	0.272	0.290	0.305	0.307	0.571	0.582	0.585	0.593	0.665	0.676	0.687	0.692
VHC	0.271	0.368	0.351	0.369	0.627	0.632	0.644	0.656	0.766	0.772	0.778	0.779

Table 2: MAP text retrieval image txt2img.

	WIKI data set				NUS-WIDE data set				MIRFlickr data set			
	16	32	64	128	16	32	64	128	16	32	64	128
CCA	0.168	0.159	0.154	0.150	0.371	0.365	0.362	0.360	0.579	0.574	0.572	0.570
SCM	0.226	0.246	0.249	0.253	0.535	0.540	0.542	0.539	0.615	0.624	0.628	0.631
SEPH	0.631	0.658	0.659	0.669	0.683	0.695	0.693	0.708	0.710	0.744	0.727	0.744
GSPH	0.645	0.663	0.671	0.674	0.681	0.697	0.686	0.714	0.726	0.742	0.748	0.764
VHC	0.487	0.748	0.751	0.757	0.686	0.715	0.761	0.776	0.766	0.780	0787	0.791

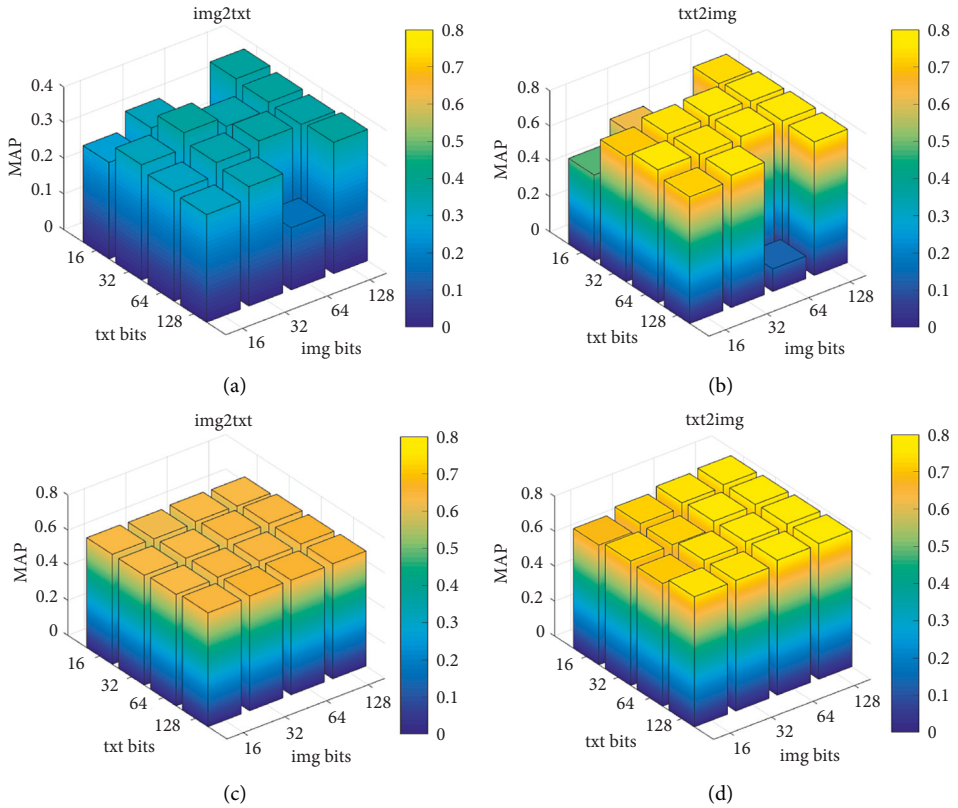
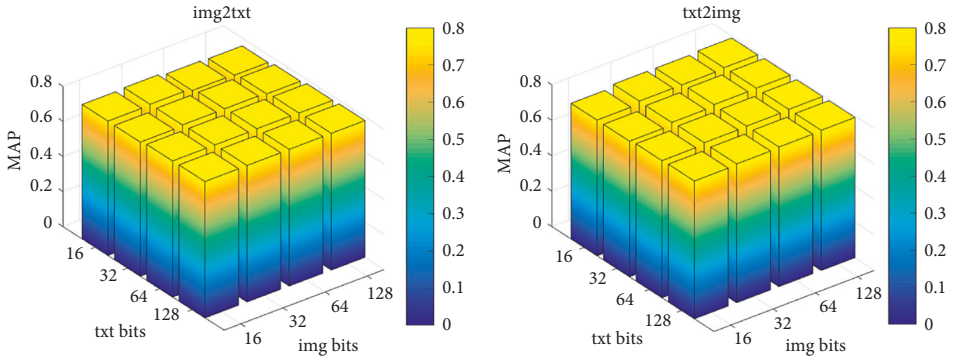


Figure 3: Continued.



(e) (f)

Figure 3: 3-D histogram of MAP index of different hash coding length combinations: (a) WIKI (img2txt), (b) WIKI (txt2img), (c) NUS-WIDE (img2txt), (d) NUS-WIDE (img2txt), (e) MIRFlickr (txt2img), and (f) MIRFlickr (txt2img).

paper (image hash coding length ∗ text hash coding length). To show the variation tendency of different hash length combinations, the curve colors of hash coding length combinations from 16 ∗ 16 to 128 ∗ 128 gradually change from dark blue, light blue, light red, and then dark red as shown in Figure 2. Generally speaking, with the growth of image hash coding, the cross-modal retrieval effect also becomes better, especially for the subgraphs (d) and (*f*) of Figure 2. In addition, Figure 2 also shows that the cross-modal retrieval model of variable-length hash coding in this paper has a more significant impact on WIKI data sets.

From the MAP three-dimensional histogram in Figure 3, it can be seen that the same and fixed hash code length cannot be set for all datasets. To be special, the optimal hash code combination is 48 ∗ 64(text ∗ image) for the img2txt task on the NUS-WIDE dataset. But the optimal hash code length combination is 32 ∗ 64 (text ∗ image) for the img2txt task on the MIRFlickr dataset to implement the img2txt task. The reason is that the text information of NUS-WIDE is richer and more hash codes are needed to represent text features. From another point of view, for some retrieval tasks, using a shorter hash code length can also achieve a comparable retrieval effect. Thus, we can conclude that using a variable-length hash code can balance the data redundancy and retrieval accuracy.

5. Conclusion

In this paper, a variable-length hash coding-based cross-modal retrieval algorithm is first proposed, which projects multimodal data into the optimal hash length space of each modal data. The similarity matrix of multimodal data is constructed according to the label matrix of each mode, and the semantic similarity relationship of the original data is still guaranteed after the multimodal hash coding is projected into the potential abstract semantic space. Then the binary constraint condition of the hash coding is always maintained in the process of optimizing the model, so that the learned multimode hash coding can better represent the original multimode data. A wide variety of experiments on WIKI datasets, NUS-WIDE datasets, and MIRFlickr datasets show that the performance of the proposed method is generally superior to that of the correlation benchmark algorithms. Therefore, the method in this paper is feasible and effective. Compared with the deep learning-based hashing methods, the retrieval performance is relatively low. Thus, in our future work, we will embed the proposed similarity matrix into the deep learning-based method to further improve the retrieved accuracy and effectively measure the relationship among multiple source data.

Data Availability

The datasets used and/or analyzed during the current study are available from the author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant 61871062), Science and Technology Research Program of Chongqing Municipal Education Commission (Grants KJQN201800615 and KJQN201900609), Natural Science Foundation of Chongqing, China (Grants cstc2020jcyj-zdxmX0024 and cstc2021jcyj-msxm2025), Key Project of Science and Technology of Chongqing Municipal Education Commission (Grant KJZDK201901203), and University Innovation Research Group of Chongqing (Grants CXQT20017 and CXQT20024).

References

[1] J. Xiong, J. Ren, L. Chen et al., "Enhancing privacy and availability for data clustering in intelligent electrical service of IoT," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1530–1540, 2019.

[2] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: concepts, methodologies, benchmarks, and

challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2372–2385, 2018.

[3] J. Xiong, R. Bi, Y. Tian, X. Liu, and D. Wu, "Toward lightweight, privacy-preserving cooperative object classification for connected autonomous vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2787–2801, 2022.

[4] J. Xiong, M. Zhao, M. Z. A. Bhuiyan, L. Chen, and Y. Tian, "An AI-enabled three-party game framework for guaranteed data privacy in mobile edge crowdsensing of IoT," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 2, pp. 922–933, 2021.

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

[6] X. Fu, K. Huang, E. E. Papalexakis et al., "Efficient and distributed generalized canonical correlation analysis for big Multiview data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2304–2318, 2019.

[7] J. Costa Pereira, E. Coviello, G. Doyle et al., "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

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

[9] J. Xiong, R. Ma, L. Chen et al., "A personalized privacy protection framework for mobile crowdsensing in IIoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4231–4241, 2020.

[10] F. Taherkhani, V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Error-corrected margin-based deep cross-modal hashing for facial image retrieval," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 3, pp. 279–293, Jul. 2020.

[11] Z. D. Chen, C. X. Li, X. Luo, L. Nie, W. Zhang, and X. S. Xu, "SCRATCH: a scalable discrete matrix factorization hashing framework for cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2262–2275, Jul, 2020.

[12] D. Wang, X. Gao, X. Wang, L. He, and B. Yuan, "Multimodal discriminative binary embedding for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4540–4554, 2016.

[13] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[14] D. Mandal, K. N. Chaudhury, and S. Biswas, "Generalized semantic preserving hashing for N-label cross-modal retrieval," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2641, Honolulu, Hi, USA, July 2017.

[15] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao, "A fast optimization method for general binary code learning," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5610–5621, Dec, 2016.

[16] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Fast supervised discrete hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 490–496, 2018.

[17] J. Xiong, X. Chen, Q. Yang, L. Chen, and Z. Yao, "A task-oriented user selection incentive mechanism in edge-aided mobile crowdsensing," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2347–2360, 2020.

[18] T. S. Chua, J. Tang, and R. Hong, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of the 2009 Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, ACM, 2009.

[19] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, ACM, Santorini Island, Greece, July 2008.