Hindawi
Mobile Information Systems
Volume 2020, Article ID 4164692, 17 pages
https://doi.org/10.1155/2020/4164692

Research Article

Modality-Dependent Cross-Modal Retrieval Based on Graph Regularization

Guanhua Wang, Hua Ji, Dexin Kong, and Na Zhang

School of Information Science and Engineering, Shandong Normal University, Jinan 250358, Shandong, China

Correspondence should be addressed to Hua Ji; jihua@sdnu.edu.cn

Received 10 October 2019; Revised 12 December 2019; Accepted 3 January 2020; Published 31 January 2020

Academic Editor: Massimo Condoluci

Nowadays, the heterogeneity gap of different modalities is the key problem for cross-modal retrieval. In order to overcome heterogeneity gaps, potential correlations of different modalities need to be mined. At the same time, the semantic information of class labels is used to reduce the semantic gaps between different modalities data and realize the interdependence and interoperability of heterogeneous data. In order to fully exploit the potential correlation of different modalities, we propose a cross-modal retrieval framework based on graph regularization and modality dependence (GRMD). Firstly, considering the potential feature correlation and semantic correlation, different projection matrices are learned for different retrieval tasks, such as image query text (I2T) or text query image (T2I). Secondly, utilizing the internal structure of original feature space constructs an adjacent graph with semantic information constraints which can make different labels of heterogeneous data closer to the corresponding semantic information. The experimental results on three widely used datasets demonstrate the effectiveness of our method.

1. Introduction

With the rapid growth of multimedia information, the representing form of information becomes rich day by day in the era of big data. The ways people obtained information have also evolved to include newspapers, websites, Weibo, and WeChat. The rapid development of mobile network provides a convenient resource platform for people. People can search a lot of information by using search engines of various websites on mobile devices according to their own needs. The structures of modal data which can be used in the mobile network are various, making it difficult to display the information needed in mobile devices accurately. Most of the retrieval methods, such as text [1–3], image [4–7], and video [8–11] retrieval, focus on single-modality retrieval [12–15], in which the query sample and retrieve sample must be performed on the same data type. Nowadays, the same thing can be expressed in different ways, and there is a growing demand for diversified forms of information expression. For example, when tourists are sightseeing, they record a wonderful journey by taking photos or recording videos. These photos and videos present the same range of content although they represent different types of media objects. Similarly, information about singers and album images is used to search for the corresponding songs, so as to obtain more information about the songs. People retrieve image data or video data related to its semantic information through text data, but different dimensions and attributes of multimedia data lead to obvious feature heterogeneity between different modalities. So the practical application of large-scale data similarity retrieval needs more effective solutions. To solve this problem, the features of different modal data need to be extracted effectively, and the retrieval method is used effectively to get more accurate information in a large amount of information.

To solve the heterogeneous problem of cross-modal retrieval [16–20], subspace learning methods have been proposed. Although different modalities have different original feature spaces, we can project such modalities into a common potential space [21]. Specifically, the most traditional feature learning method called canonical correlation analysis (CCA) [22] maximized the correlation between two couples of different modalities' features and obtained low-dimensional expressions with high correlations of different modalities in a common potential space. CCA is a simple algorithm for realizing the feature space association. Based on CCA, Hwang et al. proposed kernel canonical correlation analysis (KCCA) [19], which obtains the correlation between image and text through cross-view retrieval in a high-dimensional feature space. The partial least squares (PLS) [23] method measured the similarity between different modalities through visual feature space to text feature space. The potential correlation of the cross-modal data obtained by the above methods through linear projection is limited, and it cannot effectively improve the performance of cross-modal retrieval. The unsupervised cross-media retrieval method only obtains pairwise information of different modalities during the subspace learning process without obtaining accurate information of high-level semantics. Another method called T-V CCA [20] obtained high-level semantics by considering the semantic class view as the third view. The correlation between different modalities is enhanced by learning semantic information. Therefore, the linear regression term is applied to a cross-modal retrieval framework, and the semantic structure is maintained. So the regression error of different modalities data is minimized.

The deep learning method has a strong ability of nonlinear learning. The deep canonical correlation analysis (DCCA) [24] combines DNN and CCA to learn more complex nonlinear transformation between different modalities data. Peng et al. proposed cross-media multiple deep networks (CMDNs) [25], which use hierarchical structures to hierarchically combine independent representations of different modalities. In addition, Wei et al. proposed deep semantic matching (deep-SM) [26] to use the CNN feature for deep semantic matching to improve the retrieve accuracy. The above method makes use of the neural network to measure the similarity of different modal data well but ignores the similarity within single modality and the similarity between the modalities. The complex latent correlations of different modalities data can be well learned by using graph regularization. The application of graph regularization [27, 28] in cross-modal retrieval lies in the construction of the graph model, maintaining the similarity between the projected data through the edges of the graph model. The graph regularization not only enhances semantic relevance but also learns intramodality and intermodality similarity. The cross-modal retrieval models we have mentioned are learned through joint distribution in a common space. On the basis of subspace learning, the correlation between multimodal data is further mined to improve the performance of cross-media retrieval.

In this paper, we propose a cross-modal retrieval framework (Figure 1) based on graph regularization and modality dependence (GRMD). The method measures the distances between different modalities' projection matrices in the semantic subspace and obtains the similarity of different modalities. The projection matrices of different modalities belonging to the same label should be as similar as possible. In the process of feature mapping, two different projection matrices are mapped into their respective semantic spaces through two linear regressions. Correlation analysis can project original data into a potential subspace, and multimodal data of the same labels can be correlated.

The main advantages of our method can be summarized as follows:

(i) The construction of the label graph enhances the consistency of the internal structure of the heterogeneous data feature space and the semantic space. The graph model of different modal data is constructed for different retrieval tasks, which not only maintains the similarity between different modal data after projection but also deepens the correlation between multimodal data and corresponding semantic information.

(ii) Heterogeneous data are projected into the semantic space of different modalities in different retrieval tasks. In different cross-modal tasks learning, different transformation matrices are obtained by combining semantic correlation and feature clustering. The mapping of media data of different modalities is achieved from the underlying features to high-level semantics, and the accuracy of subspace learning is improved by using semantic information. This approach not only retains the similarity relationship of multimodal samples but also makes the semantic information more accurately understood in the projection process.

(iii) The results of experiments that we carried out on three datasets indicate that the proposed framework is superior to other advanced methods.

2. Related Work

We briefly introduce several related methods in this section. Most cross-modal retrieval methods focus on joint modeling of different modalities. Image and text retrieval are the main subjects of cross-modal retrieval research. The representation features of different modalities are not only inconsistent but also located in different feature spaces. By learning potential common subspaces, data of different modalities are mapped to common isomorphic subspaces for retrieval from traditional heterogeneous spaces.

Subspace learning plays an important role in cross-modal problems, and the most traditional unsupervised method is canonical correlation analysis (CCA) [22], which maps heterogeneous data into isomorphic subspaces, maximizing the correlation of the two couples of features. It only uses the information of the multimodal pair, ignoring the importance of labels' information, and the result of the search is not optimal. Heterogeneous data with the same semantics are interrelated in a common semantic space. After the data have been projected into the isomorphic feature space, the supervised method (SCM) [22], which combines CCA and SM, generates a common semantic space for CCA representation learning by linear regression to improve retrieval performance. In addition to CCA, Sharma et al. proposed a generalized multiview analysis (GMA) [29]

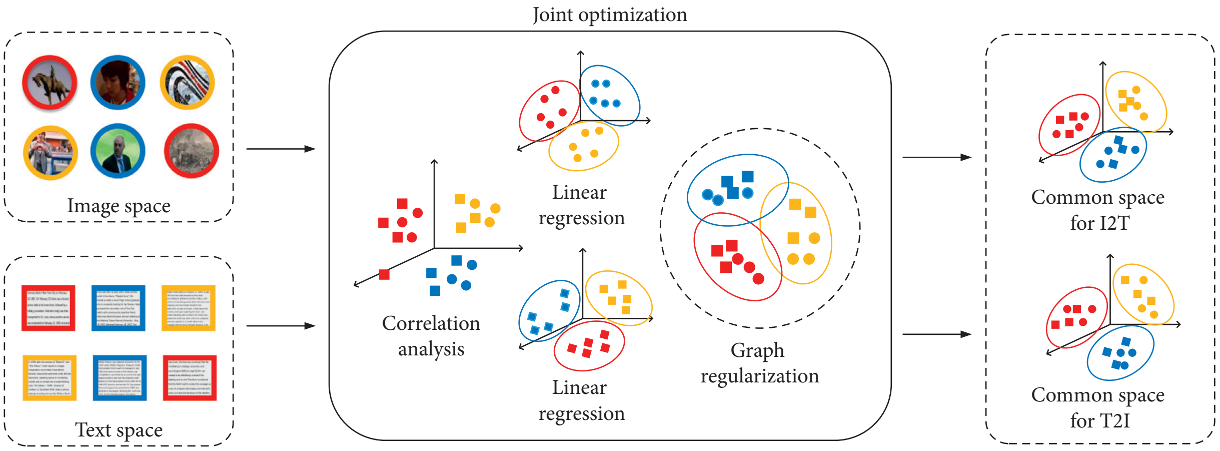


Figure 1: Flowchart of our proposed method.

for learning a common subspace through a supervised extension of CCA for cross-modal retrieval.

It is limited to improve the retrieval performance by learning the potential relationship between different modalities data. The retrieval method [30] based on deep learning can better combine the feature extraction of samples with the learning of common space, to obtain better retrieval result. Andrew et al. proposed deep canonical correlation analysis (DCCA) [24] nonlinear learning of CCA to learn complex nonlinear transformations of different modalities, through the corresponding constraints of the corresponding subnetworks to make data highly linearly related. Srivastava et al. proposed deep Boltzmann machines (DBMs) [31], which is an algorithm that learns generalization models, and thus enhances the effectiveness of retrieval. In addition, other deep models are used for cross-modal retrieval by exploiting the relevance of enhanced multimedia data. Peng et al. [32] proposed constructing a multipathway network, using coarse-grained instances and fine-grained patches to improve cross-modal correlation and achieve the best performance. The cross-modal retrieval method based on DNN uses DNN to learn the nonlinear relationship of different modalities, and the training data play a key role in the learning process. In [33], Huang et al. proposed the modal-adversarial hybrid transfer network (MHTN), an end-to-end architecture with a modal-sharing knowledge transfer subnetwork, and a modal-adversarial semantic learning subnetwork. It enhances the semantic consistency of the data, making the different modalities aligned with each other. Yu et al. proposed the graph in network (GIN) [34], which learns text representation to get more semantically related words through the graph convolution network. In the learning process, the semantic information is promoted significantly; the data information is extracted effectively; and the retrieval accuracy is improved better.

In addition, different feature representations of different modalities data cause the problem that cross-modal data cannot be effectively established. The uniform sparse representations of different modalities data are obtained through dictionary learning, but accurate semantic relationships cannot be obtained through dictionary learning alone. Semantic differences are reduced by using semantic constraints. Therefore, semantic differences should be reduced through semantic constraint methods. Semantic information is used to project sparse representations of different modalities in the semantic space to perform cross-modal matching for more accurate understanding and retrieval. A dictionary learning algorithm [35, 36] proposed by Xu et al. uses the learning of a coupled dictionary to update the dictionary that optimizes different modalities and obtains the sparse representation corresponding to different modalities data. With the rapidly increasing availability of high-dimensional data, hash learning for cross-modal retrieval has emerged. The hash learning method not only projects high-dimensional data into Hamming space but also preserves the original structure of data features as much as possible. Multiscale correlation sequential cross-modal hashing learning (MCSCH) [37] is a multiscale feature-guided sequential hashing learning method that can mine multiscale correlations among multiscale features of different modalities. In the process of cross-modal hash learning, the correlation of similar data is maximized and the correlation of dissimilar data is minimized.

Complex correlation between different modalities cannot be fully considered, but cross-modal retrieval method [38] based on graph regularization can learn complex potential correlation of different modalities data by building graph models. The graph regularization [39] is used to maintain intrapair and interpair correlations and perform feature selection for different feature spaces. Zhai et al. proposed a joint representation learning algorithm (JGRHML) [27] to consider heterogeneous relationships in a joint graph regularization. The algorithm optimizes the correlation and complementarity of different modalities data and obtains related information between heterogeneous data through nearest neighbors. To improve the JGRHML algorithm, joint representation learning (JRL) [28] proposed by Zhai et al. maintains the structural information between the original data through k-nearest neighbors, and it added the semantic regularization term to integrate the semantic information of the original data. The cross-modal retrieval methods we have mentioned that use adjacent graphs to learn the potential space and maintaining multimodal feature correlation, simultaneously maintaining local relationship, also significantly improve the retrieval performance.

We propose a method based on modality-dependence and graph regularization. In a common semantic subspace, data with the same semantics are similar to each other through potential relationships. Wei et al. proposed a modality-dependent cross-media retrieval method [40]. The method focuses on the retrieval direction and uses the semantic information of the query modality to project the data into the semantic space of the query modality. It considers not only the direct correlation between different modalities but also the low-level features that do not combine well with the nonlinear association. Although this method cannot fully describe the complex correlation between different modalities data, inspired by this method, we can use graph regularization to further analyze the potential correlation of data. Compared with the abovementioned methods, we maintain the correlation between data structure information and semantic information by integrating modal data information into a semantic graph and learning different projection matrices and semantic spaces for different retrieval tasks. Readers can learn more about our methods from the following explanation of how we have achieved good retrieval results.

The paper is organized as follows. Section 2 briefly introduces the relevant methods of cross-modal retrieval. In Section 3, the method we propose is described in detail. Section 4 presents our experimental results and the analysis of a comparison with other methods. Section 5 concludes this paper.

3. Modality-Dependent Cross-Modal Retrieval Based on Graph Regularization

In this section, we first introduce the notation and problem definitions associated with the objective function and then propose the overall cross-modal learning framework for GRMD. Finally, an effective iterative approach is proposed to complete this framework.

3.1. Notation and Problem Definition. Let $X = [X_1, X_2, \ldots, X_n] \in R^{p\times n}$ and $Y = [Y_1, Y_2, \ldots, Y_n] \in R^{q\times n}$ denote the feature matrices of image data and text data, respectively. $S = [S_1, S_2, \ldots, S_n] \in R^{c\times n}$ represents a semantic matrix with a number of labels C. The i-th row of the semantic matrix is the semantic vector corresponding to X_i and Y_i, $S(i, j) = 1$; otherwise, $S(i, j) = 0$. The image projection matrix and the text projection matrix in I2T are represented by $U \in R^{p\times c}$ and $V \in R^{q\times c}$. The descriptions of important notations frequently used in this paper are listed in Table 1.

Table 1: Summary of notation.

Notation	Description
n	Number of training samples
S	Semantic matrix of image and text
p and q	Dimensions of image and text
$X = [X_1, X_2, \ldots, X_n] \in R^{p\times n}$	Feature matrix of image
$Y = [Y_1, Y_2, \ldots, Y_n] \in R^{q\times n}$	Feature matrix of text
$U = [U_1, U_2, \ldots, U_n] \in R^{p\times c}$	Projection matrix of image
$V = [V_1, V_2, \ldots, V_n] \in R^{q\times c}$	Projection matrix of text
λ, α, β_1, and β_2	Balance parameters

3.2. Objective Function. Our goal is to keep the semantic consistency of multimodal data in the process of mapping different patterns of data to a common potential space. In different retrieval tasks, there are three important factors, semantic information, data correlation, and data structure distribution, each of which interacts on the other two. Therefore, semantic subspace is used as a common potential space in this paper. Through the association of potential space and semantic space, semantic information enables samples of the same category to be mapped to nearby locations:

$$F(U,V) = \lambda L(U,V) + (1-\lambda)S(U,V) + \alpha H(U,V) + R(U,V), \tag{1}$$

where $F(U,V)$ consists of four terms. $L(U,V)$ is a correlation analysis term that keeps samples of the same class close to each other. $S(U,V)$ is a linear regression that maps data of different modalities into the semantic space. $H(U,V)$ is a graph regularization term that uses the modal graph to enhance the intramodal similarity. $R(U,V)$ is a regularization term that preserves the stability of projection matrices.

3.2.1. The First Term. The first term is a correlation analysis term that minimizes the difference between multimodal data in a potential subspace. Different modality data need to remain close to each other in potential subspaces. The representations of the paired heterogeneous data in the common subspace should be as similar as possible, and thus, the distance between the two should be as small as possible:

$$L(U,V) = \left\|U^T X - V^T Y\right\|_F^2. \tag{2}$$

This term reduces the distance between multimodal data of the same label, thus improving the correlation between them.

3.2.2. The Second Term. The second term is a linear regression, which transforms the feature space of query modality into semantic space. This term only considers the query modality semantic, which is more pertinent and effective than that of considering both the query modality semantics and the retrieval modality semantics. The improvement in the accuracy of the mapping of query modality data can ensure the accuracy of subsequent retrieval. Once the label of the query modality data has been incorrectly predicted, it is difficult to ensure that other related modalities data are retrieved in subsequent steps:

$$S(U,V) = \left\|U^T X - S\right\|_F^2, \tag{3}$$

$$S(U,V) = \left\|V^T Y - S\right\|_F^2. \tag{4}$$

This term focuses on the differences between different retrieval tasks and learns two different projection matrices for different retrieval tasks. It transforms the query modality data from the original feature space into the corresponding semantic space, and similar data are centrally distributed in the semantic subspace.

3.2.3. The Third Term. Here, we preserve the original distribution of different modalities data in the common subspace as much as possible by adding a graph regularization term in the objective function. The neighboring data points are as close as possible to each other in the common subspace. We define an undirected symmetric graph $H = (V_x, W_x)$, where V_x is the set of data in X and W_x is the similarity matrix. Element W_{ij} of W_x is defined as follows:

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \in N_k,\left(X_j\right) \text{ or } x_j \in N_k\left(X_i\right) \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $N_k(X_j)$ represents k neighbors of X_j that are obtained by calculating the distance between data pairs in the original space and selecting the nearest k neighbors.

$$L = E - D^{-1/2} W D^{-1/2}, \tag{6}$$

where L is a symmetric semidefinite matrix, D is a diagonal matrix, and the diagonal elements are $d_{ii} = \sum_j w_{ij}$.

By constructing a local label graph for each modality through semantic information, the structure of the feature space can be made consistent with that of the label space. In the shift between different modalities, the internal structure of modalities is preserved so that different modalities data in the same label should as near as possible after mapping:

$$\begin{aligned} H\left(U_1, V_1\right) &= \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \left\| \frac{U_1^T X_i}{\sqrt{d_{ii}}} - \frac{U_1^T X_j}{\sqrt{d_{jj}}} \right\|_2^2 \\ &\quad - \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \left\| \frac{S_i}{\sqrt{d_{ii}}} - \frac{S_j}{\sqrt{d_{jj}}} \right\|_2^2 \\ &= tr\left(U_1 X^T L_1 X U_1^T - S^T L_1 S\right). \end{aligned} \tag{7}$$

Similarly, we calculate the similarity matrix W, the symmetric matrix D, and the Laplacian matrix L of the text, and the regularization terms of the text are defined as follows:

$$H\left(U_2, V_2\right) = tr\left(V_2 Y^T L_2 Y V_2^T - S^T L_2 S\right). \tag{8}$$

3.2.4. The Forth Term. The fourth term is the regularization term that controls the complexity of the projection matrix and prevents overfitting. Therefore, the constraints of the term can control the stability of the obtained values. Parameters β_1 and β_2 balance the regularization term:

$$R(U,V) = \beta_1 \|U\|_F^2 + \beta_2 \|V\|_F^2. \tag{9}$$

For I2T:

The algorithm we present learns a pair of projection matrices U_1 and V_1 through the image query text (I2T), and our final objective function is specifically expressed as follows:

$$\begin{aligned} F\left(U_1, V_1\right) &= \lambda \left\|U_1^T X - V_1^T Y\right\|_F^2 + (1-\lambda)\left\|U_1^T X - S\right\|_F^2 \\ &\quad + \alpha tr\left(U_1 X^T L_1 X U_1^T - S^T L_1 S\right) + \beta_1 \left\|U_1\right\|_F^2 \\ &\quad + \beta_2 \left\|V_1\right\|_F^2. \end{aligned} \tag{10}$$

For T2I:

Similarly, the objective function of T2I is expressed as follows:

$$\begin{aligned} F\left(U_2, V_2\right) &= \lambda \left\|U_2^T X - V_2^T Y\right\|_F^2 + (1-\lambda)\left\|V_2^T Y - S\right\|_F^2 \\ &\quad + \alpha tr\left(V_2 Y^T L_2 Y V_2^T - S^T L_2 S\right) \\ &\quad + \beta_1 \left\|U_2\right\|_F^2 + \beta_2 \left\|V_2\right\|_F^2. \end{aligned} \tag{11}$$

As expressed by (11), a cross-modal retrieval problem retrieves related image modalities based on the text modality. In contrast to (3), our linear regression term is a text feature space conversion to a semantic text space, rather than a semantic image space in I2T. The image projection matrix and the text projection matrix in T2I are represented by $U_2 \in R^{c \times p}$ and $V_2 \in R^{c \times q}$.

3.3. Iterative Optimization for the Proposed Algorithm. In this section, both (10) and (11) are nonconvex optimization problems, so we design an algorithm to find fixed points. We observe that if another item is fixed, equation (10) is convex to the other item. Similarly, equation (11) is fixed while the other item is fixed, and the other item is also convex. Therefore, by using the gradient descent method, we can achieve the minimization of the other term by fixing one of $U_1\left(U_2\right)$ or $V_1\left(V_2\right)$.

First, we compute the partial derivative of $F\left(U_1, V_1\right)$ with respect to U_1 and set it to 0:

$$\begin{aligned} \frac{\partial F\left(U_1 V_1\right)}{\partial U_1} &= \lambda\left(X X^T U_1 - X Y^T V_1\right) + (1-\lambda)\left(X X^T U_1 - X S^T\right) \\ &\quad + \beta_1 U_1 + \alpha X^T L_1 X U_1. \end{aligned} \tag{12}$$

Similarly, we compute the partial derivative of $F\left(U_1 V_1\right)$ with respect to V_1 and set it to 0:

$$\frac{\partial F\left(U_1 V_1\right)}{\partial V_1} = \lambda\left(Y Y^T V_1 - Y X^T U_1\right) + \beta_2 V_1. \tag{13}$$

According to the above formula, the resulting solutions are, respectively, as follows:

$$U_1 = \left(XX^T + \beta_1 I + \alpha X^T L_1 X\right)^{-1} \cdot \left(XS^T + \lambda XY^T V_1 - \lambda XS^T\right), \tag{14}$$

$$V_1 = \lambda YX^T U_1 \left(\lambda YY^T + \beta_2 I\right)^{-1}. \tag{15}$$

Similarly, for T2I, $F(U_2 V_2)$ is biased for U_2 and V_2, respectively. U_2 and V_2 are updated iteratively until the results converge:

$$\begin{aligned} U_2 &= \lambda XY^T V_2 \left(\lambda XX^T + \beta_1 I\right)^{-1}, \\ V_2 &= \left(YY^T + \beta_2 I + \alpha Y^T L_2 Y\right)^{-1} \left(YS^T + \lambda YX^T V_2 - \lambda YS^T\right). \end{aligned} \tag{16}$$

The main optimization procedure of the method we present for I2T is given in Algorithm 1, and the T2I task is similar to the I2T task.

4. Experiments

The methods we present in this section are tested experimentally on three datasets. We evaluate our proposed method by comparison with other advanced methods.

4.1. Datasets. Three datasets detailed below are chosen for the experiment.

4.1.1. Wikipedia. The Wikipedia dataset [22] consists of 2,866 different image-text pairs belonging to 10 semantic categories selected from 2,700 "feature articles." This dataset is randomly divided into a training set with 2,173 image-text pairs and a test set with 693 image-text pairs, and these two sets are marked by 10 semantic class words. Image features are represented by 4096-dimensional CNN visual features, while the representation of text features is 100-dimensional LDA text features.

4.1.2. Pascal Sentence. The Pascal sentence dataset [26] consists of 1000 image-text pairs from 20 semantic categories. In each semantic category, there are 50 image-text pairs, 30 of which are selected as training pairs, and the rest are used as test pairs for each class. We represent image features by extracting 4096-dimensional CNN visual features and represent text features by 100-dimensional LDA text features.

4.1.3. INRIA-Websearch. The INRIA-Websearch dataset [41] has 71478 image-text pairs from 353 semantic categories, formed with 14698 image-text pairs built by selecting the largest 100 categories. This dataset is randomly divided into 70% of pairs used as a training set and 30% used as a test set. Each image and text are represented by a 4096-dimensional CNN visual feature and a 1000-dimensional LDA feature, respectively.

```
Input: training image datasets
$X = [x_1, x_2, \ldots, x_n] \in R^{p\times n}$;
Training text datasets $Y = [y_1, y_2, \ldots, y_n] \in R^{q\times n}$;
Semantic sets $S = [s_1, s_2, \ldots, s_n] \in R^{n\times c}$
Balancing parameters $\lambda$, $\alpha$, $\beta_1$, $\beta_2$
Output: projection matrices $U_1$ and $V_1$.
calculate the graph Laplacian matrix $L_1$;
initialize $U_1$ and $V_1$ to be identity matrices;
repeat
    fix $V_1$ and update $U_1$ according to (14);
    fix $U_1$ and update $V_1$ according to (15);
until convergence
end for
```

Algorithm 1: Modality-dependent cross-modal retrieval based on graph regularization in I2T.

4.2. Experimental Settings. We assume that the Euclidean distance is used to compute the similarity of data features when multimedia data are projected into a common subspace. In this part, to evaluate the results of cross-modal retrieval, we consider the widely used mean average precision (MAP) [22] scores and precision recall (PR) curves. Specifically, the average precision (AP) of each query is obtained, and their average values are calculated to obtain a MAP score:

$$AP = \frac{1}{R} \sum_{R=1}^{n} \frac{R_k}{k} \times \mathrm{rel}_k, \tag{17}$$

where n is the size of the test set and R is the number of related items. Condition $\mathrm{rel}_k = 1$ means that the item with level k is relevant. Otherwise, $\mathrm{rel}_k = 0$; R_k is the number of related items in the top k returns. To evaluate the performance of the proposed GRMD retrieval method, we compare GRMD with the canonical correlation analysis (CCA) [22], kernel canonical correlation analysis (KCCA) [19], semantic matching (SM) [22], semantic correlation matching (SCM) [22], three-view canonical correlation analysis (T-V CCA) [42], generalized multiview linear discriminant analysis (GMLDA) [29], generalized multiview canonical correlation analysis (GMMFA) [29], modality-dependent cross-media retrieval (MDCR) [40], joint feature selection and subspace learning (JFSSL) [43], joint latent subspace learning and regression (JLSLR) [44], generalized semisupervised structured subspace learning (GSSSL) [45], a cross-media retrieval algorithm based on the consistency of collaborative representation (CRCMR) [46], cross-media retrieval based on linear discriminant analysis (CRLDA) [47], and cross-modal online low-rank similarity (CMOLRS) function learning method [48]. The descriptions and characteristics of the above comparison methods used in the whole experiment are summarized in Table 2.

4.3. Experimental Results. The experiment is a cross-media retrieval of two subtasks: I2T and T2I. The traditional distance metrics are used to measure the similarity of different modalities' objects. The experiment was carried out on three

Table 2: The summarization of all compared methods.

Descriptions of comparison methods	Characteristics of comparison methods
CCA is a classic subspace method that projects different modalities into a common subspace to maximize the correlation between the paired information items.	Correlation analysis Unsupervised learning
KCCA obtains the correlation between image and text through cross-view retrieval in a high-dimensional feature space.	Kernel correlation analysis Unsupervised learning
SM projects image-text pairs into the semantic space to retrieve data from different modalities.	Semantic information
SCM projects an image-text pair to the semantic space in which learning is performed by CCA. SCM uses a combination of CCA and SM to improve retrieval performance.	Correlation analysis Semantic information
GMLDA seeks the best projection direction so that the similar samples are as close as possible, and different classes of samples are as far as possible.	Generalized multiview analysis Linear discriminant analysis Semantic information
GMMFA combines semantic information, and CCA constraints to learn a common subspace through the combination of GMA and MFA.	Generalized multiview analysis Canonical correlation analysis Semantic information
MDCR performs different retrieval tasks for different query objects. Different projection matrices are learned to optimize each retrieval result.	Different retrieval tasks Correlation analysis Semantic information
JFSSL uses graph regularization to maintain similarity between intermodality and intramodality and performs feature selection for different feature spaces, thereby improving performance.	Graph regularization Semantic information
JLSLR uses label graphs to learn the latent space and maintains a high correlation of multimodality features. The local relationships are maintained when different modal features are projected onto a common space.	Graph regularization Semantic information
GSSSL learns a discriminative common subspace by combining the relevance of samples for different modalities with the semantic information.	Graph regularization Semantic information
CRCMR not only uses dictionary learning to obtain collaborative representation for multimodal data but also takes into account the same semantic information of multimodal data.	Collaborative representation Semantic information
CRLDA improves retrieval performance by considering the pairwise correlation between image features and text features and improving the discriminative characteristic of textual modality.	Different retrieval tasks Correlation analysis Semantic information Linear discriminant analysis
CMOLRS adapts the margin of hinge loss for each triple, effectively utilizes sample features and semantic information and thus achieves a low-rank bilinear similarity measurement on data.	Relative similarities Semantic information

datasets. Tables 3–5 show the experimental results of different datasets. Later, we will study the effects of different parameter settings on the performance of GRMD.

In the experiment on the Wikipedia dataset, we set various parameters as follows: for I2T, $\lambda = 0.3$, $\alpha = 0.2$, $\beta_1 = 0.8$, and $\beta_2 = 0.5$; for T2I, $\lambda = 0.4$, $\alpha = 0.1$, $\beta_1 = 1.0$, and $\beta_2 = 0.2$. MAP scores we obtained on I2T and T2I tasks are shown in Table 3. Figures 2(a) and 2(b) show the MAP scores on the Wikipedia dataset for different retrieval tasks, and Figure 2(c) shows the MAP scores for different labels as an indication of average performance. Figures 3(a) and 3(b) show the precision-recall curves for two retrieval tasks, I2T and T2I. The results show that CCA and KCCA do not use semantic information, and its retrieval performance is poor. SM only consider the semantic information and does not consider the related data. Our approach combines data correlation and semantic information to learn heterogeneous data problems so that good retrieval performance can be achieved.

In the experiment on the Pascal Sentence dataset, we set various parameters as follows: for I2T, $\lambda = 0.4$, $\alpha = 0.2$, $\beta_1 = 0.3$, and $\beta_2 = 1.0$; for T2I, $\lambda = 0.4$, $\alpha = 0.1$, $\beta_1 = 0.4$, and $\beta_2 = 0.1$. The MAP scores that we obtained on I2T tasks and T2I tasks are shown in Table 4. Figures 2(d) and 2(e) show the MAP scores on the Pascal sentence dataset for different retrieval tasks, and Figure 2(f) shows the MAP scores for different labels as an indication of average performance. Figures 3(c) and 3(d) show the precision-recall curves for two retrieval tasks, I2T and T2I. It can be concluded from the experimental results of SCM, T-V CCA, GMLDA, GMMFA, CMOLRS, and MDCR that although they all consider data correlation and semantic information, the MAP scores of MDCR are higher because it learns different semantic subspaces for different retrieval tasks. These methods do not fully understand the complex correlation of heterogeneous data. Therefore, our method is projected not only in different semantic subspaces but also the similarity between heterogeneous data projected can be well maintained by constructing adjacent graphs. The results show that our approach is necessary for considering different retrieval tasks and maintaining the similarity of heterogeneous data.

In the experiment on the INRIA-Websearch dataset, we set various parameters as follows: for I2T, $\lambda = 0.4$, $\alpha = 0.2$, $\beta_1 = 0.3$, and $\beta_2 = 1.0$; for T2I, $\lambda = 0.3$, $\alpha = 0.2$, $\beta_1 = 1.0$, and $\beta_2 = 0.1$. The MAP scores that we obtained on I2T tasks and T2I tasks are shown in Table 5. After the semantic category is increased, the retrieval performance of our method is still very good. CRLDA only considers the discriminability of text features. JFSSL, JLSLR, and GSSSL validate the validity

Table 3: MAP scores on the Wikipedia dataset.

Method	I2T	T2I	Average
CCA	0.226	0.246	0.236
KCCA	0.332	0.351	0.342
SM	0.403	0.357	0.380
SCM	0.351	0.324	0.337
T-V CCA	0.310	0.316	0.313
GMLDA	0.372	0.322	0.347
GMMFA	0.371	0.322	0.346
MDCR	0.419	0.382	0.401
JFSSL	0.392	0.381	0.387
JLSLR	0.394	0.369	0.382
GSSSL	0.413	0.376	0.395
CRCMR	0.408	0.395	0.402
CRLDA	0.425	0.388	0.407
CMOLRS	0.424	0.382	0.403
GRMD	0.438	0.399	0.419

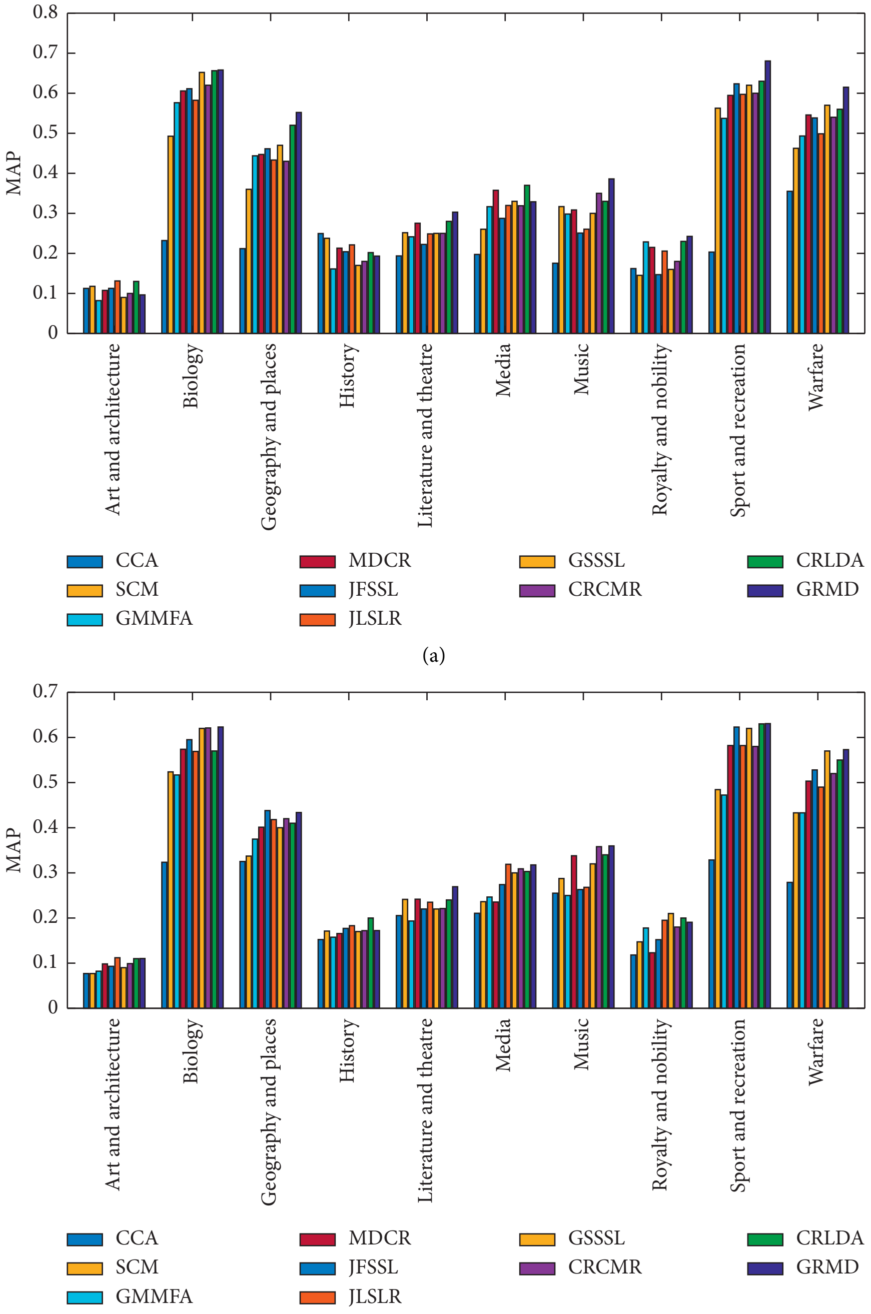
Table 4: MAP scores on the Pascal sentence dataset.

Method	I2T	T2I	Average
CCA	0.261	0.356	0.309
KCCA	0.401	0.398	0.399
SM	0.426	0.467	0.446
SCM	0.369	0.375	0.372
T-V CCA	0.337	0.439	0.388
GMLDA	0.456	0.448	0.451
GMMFA	0.455	0.447	0.452
MDCR	0.449	0.475	0.462
JFSSL	0.407	0.402	0.404
JLSLR	0.454	0.455	0.455
GSSSL	0.468	0.464	0.466
CRCMR	0.471	0.480	0.476
CRLDA	0.471	0.478	0.474
CMOLRS	0.415	0.423	0.419
GRMD	0.484	0.491	0.488

Table 5: MAP scores on the INRIA-websearch dataset.

Method	I2T	T2I	Average
CCA	0.274	0.392	0.333
KCCA	0.517	0.526	0.522
SM	0.439	0.517	0.478
SCM	0.403	0.372	0.387
T-V CCA	0.329	0.500	0.415
GMLDA	0.505	0.522	0.514
GMMFA	0.492	0.510	0.501
MDCR	0.520	0.551	0.535
JFSSL	0.533	0.562	0.548
JLSLR	0.525	0.545	0.535
GSSSL	0.530	0.552	0.541
CRCMR	0.532	0.555	0.544
CRLDA	0.531	0.552	0.542
CMOLRS	0.358	0.374	0.366
GRMD	0.539	0.558	0.549

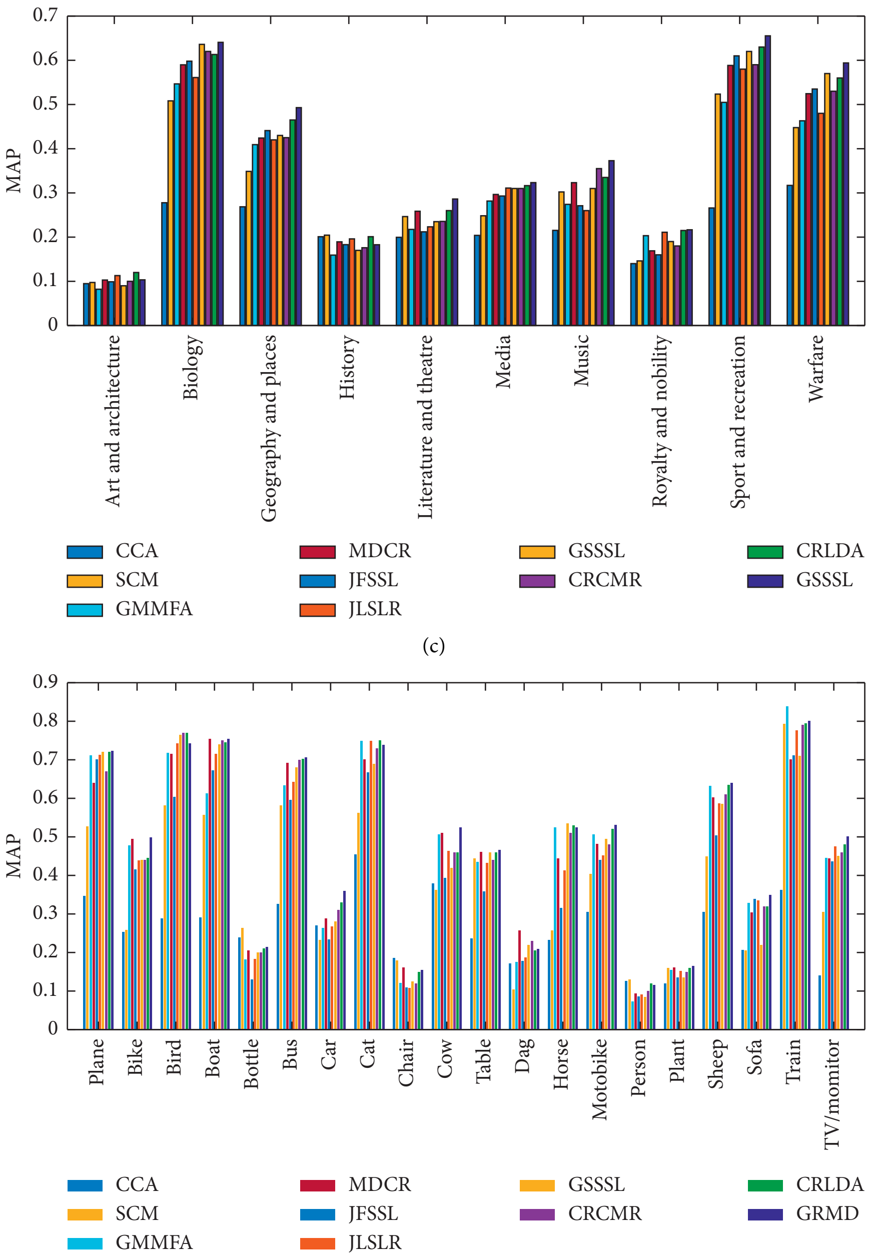
of adjacent graphs by considering the complex similarity of heterogeneous data. Our method not only considered the semantic information when constructing the adjacent graphs but also constructed the corresponding semantic graphs for different query objects. We observe that the MAP scores on T2I tasks of JFSSL is higher than that of our



(a)

(b)

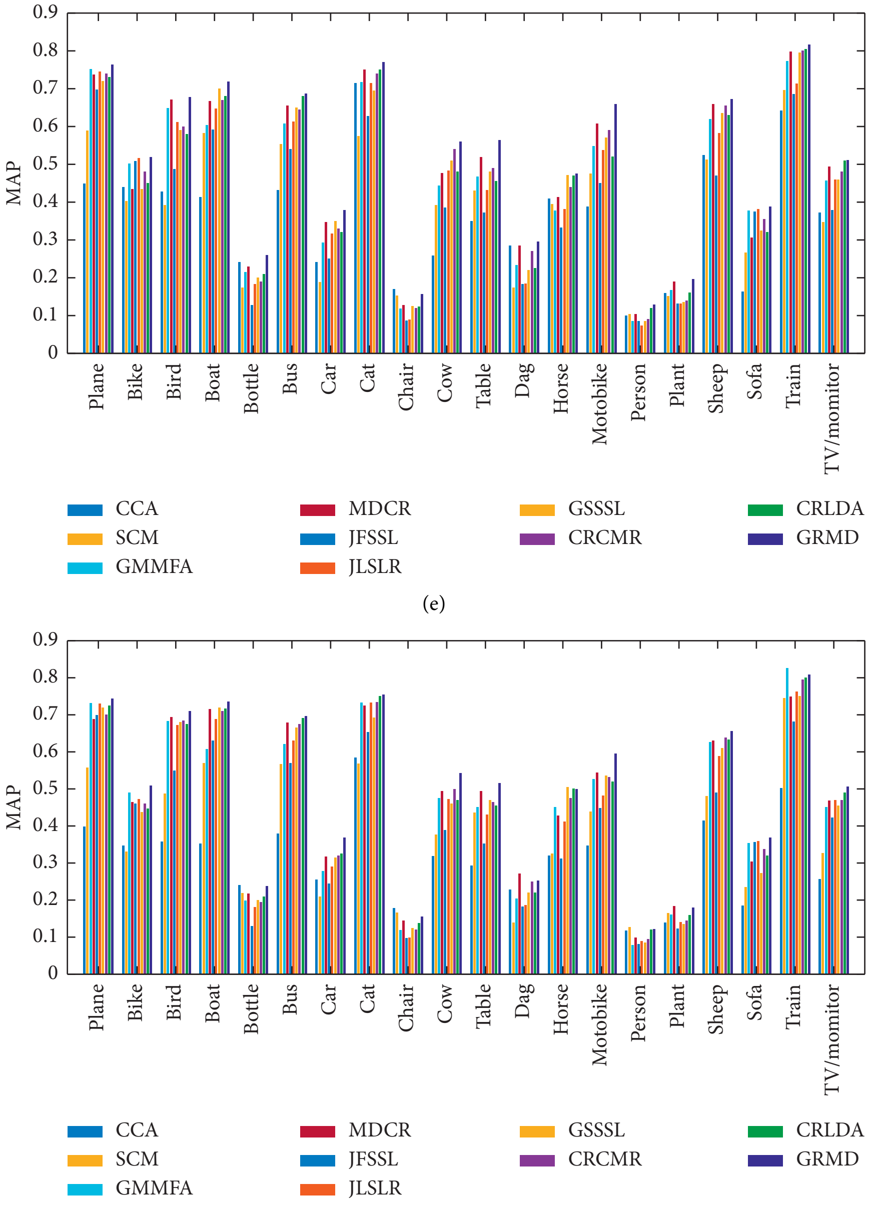
Figure 2: Continued.



(c)

(d)

Figure 2: Continued.



(e)

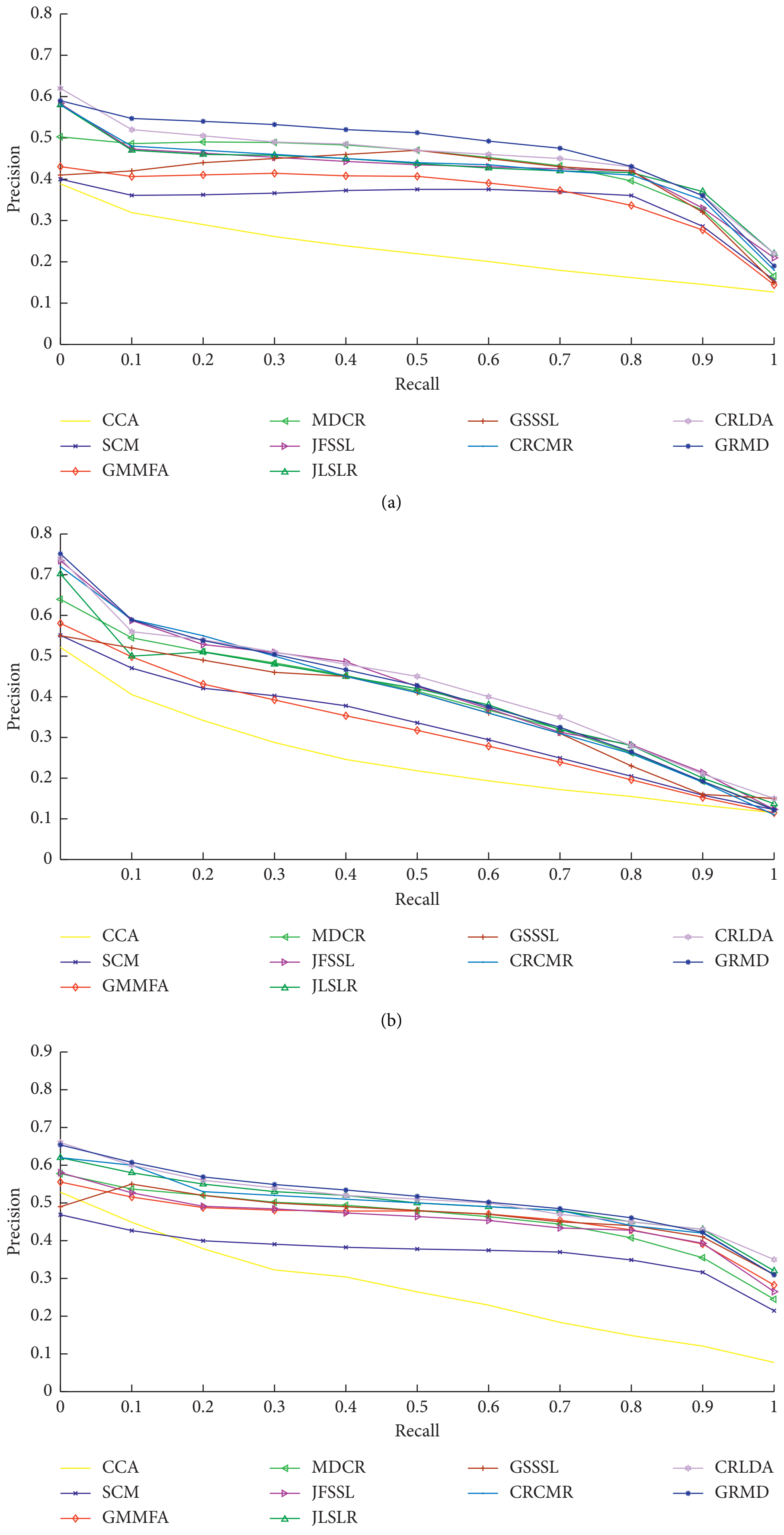
(f)

Figure 2: MAP scores for each class on two datasets: (a) I2T on the Wikipedia dataset; (b) T2I on the Wikipedia dataset; (c) average MAP on the Wikipedia dataset; (d) I2T on the Pascal sentence dataset; (e) T2I on the Pascal sentence dataset; (f) average MAP on the Pascal sentence dataset.

method. This result may be due to feature selection for heterogeneous data. Figures 3(e) and 3(f) show the precision-recall curves for two retrieval tasks I2T and T2I. A comparison with other methods shows that our method has a certain stability and performs well on retrieval tasks.

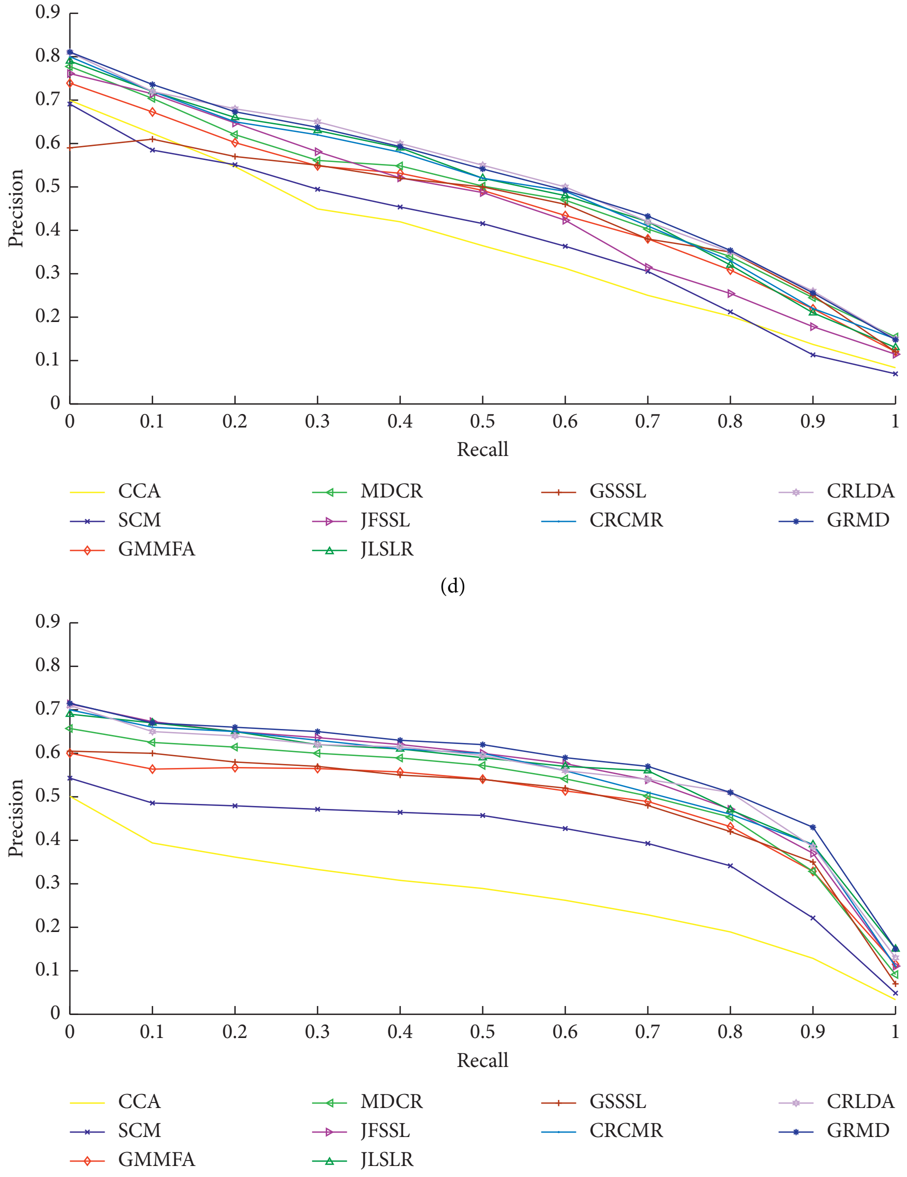
All the tables and figures below show our experimental results. We introduce two aspects of effectiveness of our method. On the one hand, the relationship between the image texts is taken into account, and only the semantics of the query object are considered. On the other hand, the semantic correlation improves retrieval precision by utilizing the local correlation of the feature map. Additionally, semantic constraints make better use of the local correlation of the feature graph and thus improve the retrieval accuracy.

4.4. Parameter Sensitivity. In this subsection, we evaluate the robustness of our approach. Our approach consists of four parameters: λ and α are balance parameters, while β_1 and β_2 are regularization parameters. In the experiment, it is observed that, with the variation in parameter λ, the retrieval performance of different retrieval tasks is stable within a wide range. Considering the results on the Pascal sentence dataset as an example, we set parameters α, β_1, and β_2 to different values during different retrieval tasks to test the sensitivity to parameter values. We tune three parameters, considering values of $\{0.0001, 0.001, 0.01, 0.1\}$. In the experiment, one parameter is fixed to observe the performance variations with other two parameters. Figures 4(a), 4(c), and 4(e) show the performance variations for I2T, and



(c)

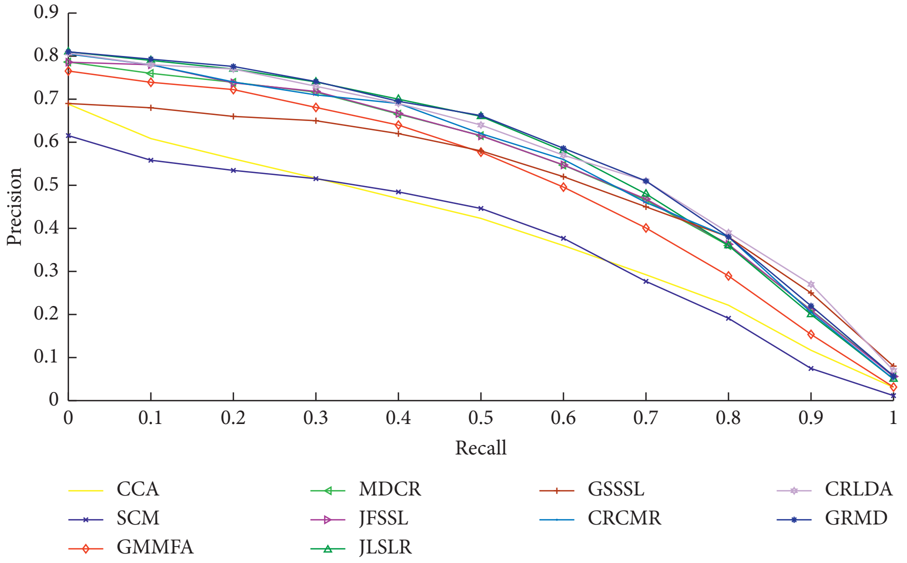
Figure 3: Continued.



(d)

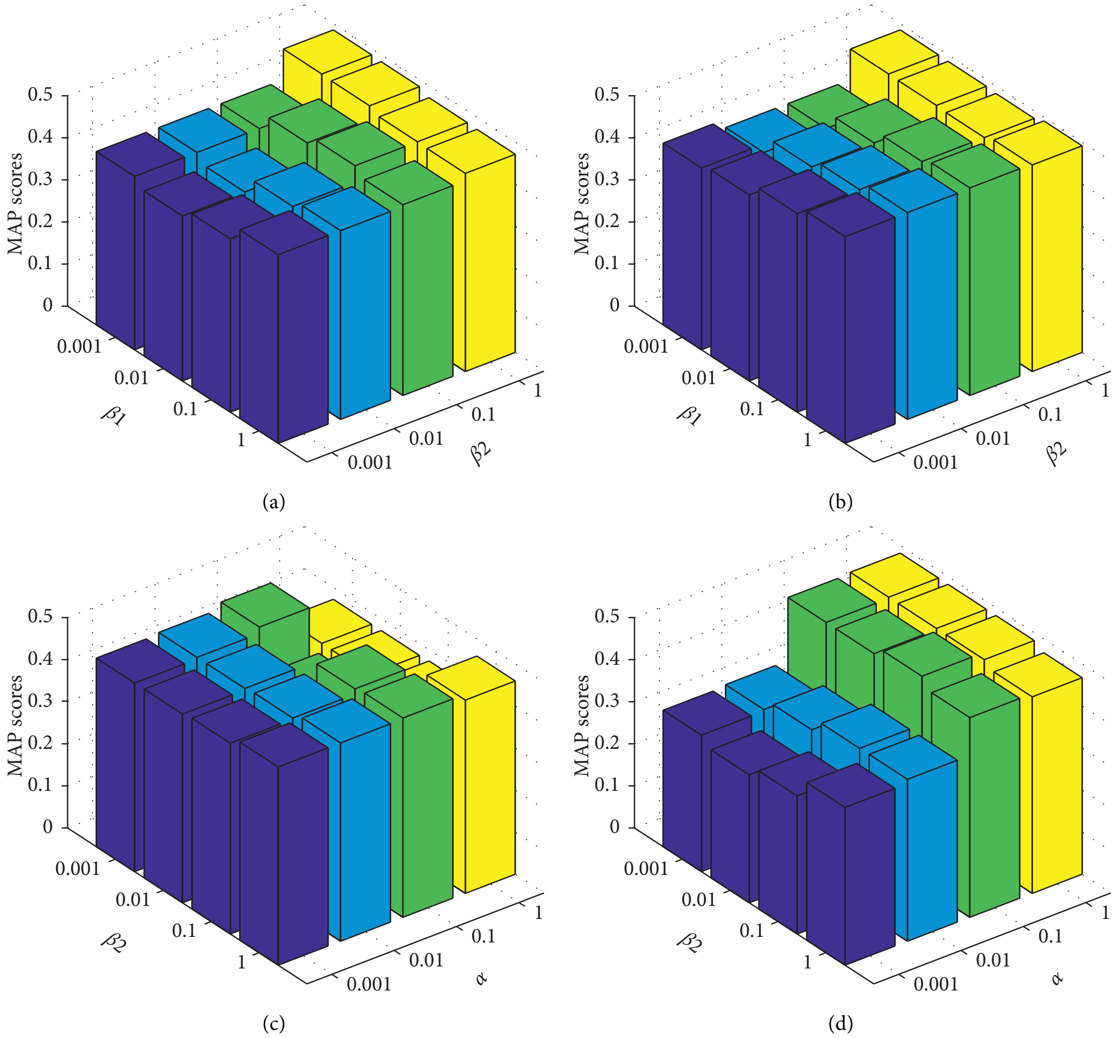
(e)

FIGURE 3: Continued.



(f)

Figure 3: Precision-recall curves of all compared methods on three datasets: (a) I2T on the Wikipedia dataset; (b) T2I on the Wikipedia dataset; (c) I2T on the Pascal sentence dataset; (d) T2I on the Pascal sentence dataset; (e) I2T on the INRIA-Websearch dataset; (f) T2I on the INRIA-Websearch dataset.



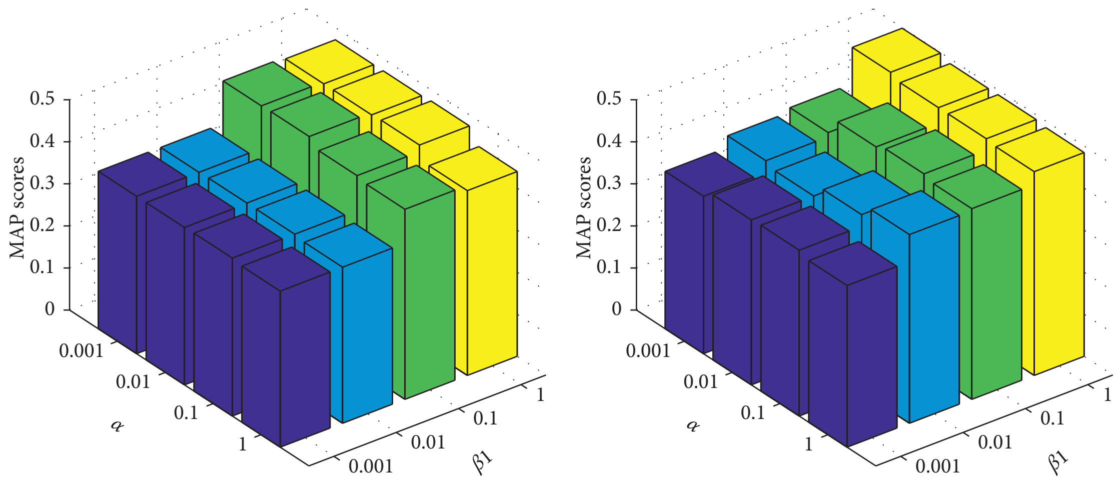
(a)

(b)

(c)

(d)

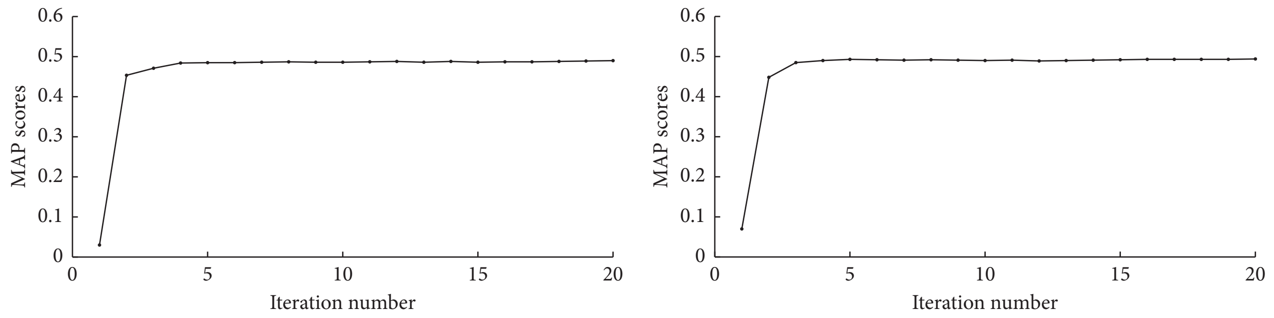
Figure 4: Continued.



(e)

(f)

FIGURE 4: Performance variations with the key parameters on the Pascal sentence dataset: (a) $\alpha = 0.1$; (b) $\alpha = 0.1$; (c) $\beta_1 = 1$; (d) $\beta_1 = 1$; (e) $\beta_2 = 1$; (f) $\beta_2 = 1$.



(a)

(b)

FIGURE 5: Convergence curves during iterations for the Pascal sentence dataset: (a) I2T; (b) T2I.

Figures 4(b), 4(d), and 4(f) show the performance variations for T2I. The figures show that our method is insensitive to these three parameters, and its performance is relatively stable.

4.5. Convergence Experiment. In this subsection, we propose an iterative optimization approach for the objective function. It is important to test its convergence during iterations. Figures 5(a) and 5(b) show convergence curves for the Pascal sentence dataset for I2T and T2I, respectively. The corresponding MAP scores tend to be stable as the number of iterations increases. The proposed approach can achieve nearly stable values within approximately seven iterations. Therefore, our approach can converge effectively and offers a stable performance.

4.6. Ablation Experiment. In Table 6, method “A” removes the graph regularization term in our approach. It means that the method uses only correlation analysis and linear regression for the features of image data and text data. The samples of different modalities are mapped to a common semantic subspace so that the multimodal data with the same label can be aggregated. Method “B” removes the correlation analysis term in our approach. It means that the paired data without a sufficient consideration of the same label should be close in a potential space. This method maintains the internal structure information of heterogeneous features.

TABLE 6: MAP scores in ablation experiments.

Dataset	Method	I2T	T2I	Average
Wikipedia dataset	Our approach	**0.438**	**0.399**	**0.419**
	A	0.409	0.374	0.391
	B	0.398	0.364	0.381
Pascal sentence dataset	Our approach	**0.484**	**0.491**	**0.488**
	A	0.445	0.456	0.451
	B	0.403	0.405	0.404
INRIA-websearch dataset	Our approach	**0.539**	**0.558**	**0.549**
	A	0.476	0.495	0.486
	B	0.436	0.510	0.473

The experimental results show the effectiveness of our method. First, to determine the corresponding projection, the data of different modalities are correlated by using the correlation between such modalities. Second, the construction of label graphs can preserve the internal structural information of the original data very well. The heterogeneous features of multimodal data are projected into a common subspace, and the multimodal data of the same label are aggregated.

5. Conclusions

In this paper, we propose a cross-modal retrieval method based on graph regularization (GRMD). This method combines the internal structure of feature space and semantic space to construct label graphs of heterogeneous data, which makes the features of different modalities closer to real labels, thus enriching the semantic information of similar data features. In addition, our method learns different projection matrices for different query tasks and also takes into account the feature correlation and semantic correlation between isomorphic and heterogeneous data features. The experimental results show that GRMD performs better than other advanced methods for cross-modal retrieval tasks. In the future, we devote to focus on the local and global structure of heterogeneous data feature distribution and to improve the retrieval framework continuously.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (grant nos. 61772322, 61572298, 61702310, and 61873151).

References

[1] A. Moffat and J. Zobel, “Self-indexing inverted files for fast text retrieval,” *ACM Transactions on Information Systems*, vol. 14, no. 4, pp. 349–379, 1996.

[2] S. Haiduc, G. Bavota, A. Marcus et al., “Automatic query reformulations for text retrieval in software engineering,” in *Proceedings of the 2013 International Conference on Software Engineering*, pp. 842–851, IEEE Press, San Francisco, CA, USA, May 2013.

[3] S. Shehata, F. Karray, and M. S. Kamel, “An efficient concept-based retrieval model for enhancing text retrieval quality,” *Knowledge and Information Systems*, vol. 35, no. 2, pp. 411–434, 2013.

[4] S. Clinchant, J. Ah-Pine, and G. Csurka, “Semantic combination of textual and visual information in multimedia retrieval,” in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, p. 44, ACM, Vancouver, Canada, October 2011.

[5] J. Yu and Q. Tian, “Semantic subspace projection and its applications in image retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 544–548, 2008.

[6] H. J. Escalante, C. A. Hérnadez, L. E. Sucar et al., “Late fusion of heterogeneous methods for multimedia image retrieval,” in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 172–179, ACM, Vancouver, Canada, October 2008.

[7] G.-H. Liu and J.-Y. Yang, “Content-based image retrieval using color difference histogram,” *Pattern Recognition*, vol. 46, no. 1, pp. 188–198, 2013.

[8] H. Greenspan, J. Goldberger, and A. Mayer, “Probabilistic space-time video modeling via piecewise GMM,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 384–396, 2004.

[9] Y. Peng and C. W. Ngo, “Clip-based similarity measure for query-dependent clip retrieval and video summarization,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 5, pp. 612–627, 2006.

[10] B. Andr, T. Vercauteren, A. M. Buchner et al., “Learning semantic and visual similarity for endomicroscopy video retrieval,” *IEEE Transactions on Medical Imaging*, vol. 31, no. 6, pp. 1276–1288, 2012.

[11] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, “Semantic pooling for complex event analysis in untrimmed videos,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1617–1632, 2017.

[12] W. Song, Y. Cui, and Z. Peng, “A full-text retrieval algorithm for encrypted data in cloud storage applications,” in *Natural Language Processing and Chinese Computing*, pp. 229–241, Springer, Berlin, Germany, 2015.

[13] M. Singha and K. Hemachandran, “Content based image retrieval using color and texture,” *Signal & Image Processing: An International Journal*, vol. 3, no. 1, pp. 39–57, 2012.

[14] X. Nie, Y. Yin, J. Sun, J. Liu, and C. Cui, “Comprehensive feature-based robust video fingerprinting using tensor model,” *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 785–796, 2017.

[15] J. Sun, X. Liu, W. Wan, J. Li, D. Zhao, and H. Zhang, “Video hashing based on appearance and attention features fusion via DBN,” *Neurocomputing*, vol. 213, pp. 84–94, 2016.

[16] Y. T. Zhuang, Y. F. Wang, F. Wu et al., “Supervised coupled dictionary learning with group structures for multi-modal retrieval,” in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, Bellevue, WA, USA, July 2013.

[17] Y. Yang, D. Xu, F. Nie, F. Luo, and Y. Zhuang, “Ranking with local regression and global alignment for cross media retrieval,” in *Proceedings of the 17th ACM International Conference on Multimedia*, pp. 175–184, ACM, Beijing, China, October 2009.

[18] Y.-T. Zhuang, Y. Yang, and F. Wu, “Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval,” *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, 2008.

[19] S. J. Hwang and K. Grauman, “Accounting for the relative importance of objects in image retrieval,” in *Proceedings of the British Machine Vision Conference*, vol. 1, no. 2, p. 5, Aberystwyth, UK, August 2010.

[20] Z. Lu and Y. Peng, “Unified constraint propagation on multi-view data,” in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, Bellevue, WA, USA, July 2013.

[21] X. Dong, E. Yu, M. Gao et al., “Semi-supervised distance consistent cross-modal retrieval,” in *Proceedings of the Workshop on Visual Analysis in Smart and Connected*

Communities, pp. 25–31, ACM, Mountain View, CA, USA, October 2017.

[22] N. Rasiwasia, J. Costa Pereira, E. Coviello et al., "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260, ACM, Firenze, Italy, October 2010.

[23] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Proceedings of the International Statistical and Optimization Perspectives Workshop "Subspace, Latent Structure and Feature Selection"*, pp. 34–51, Springer, Bohinj, Slovenia, February 2005.

[24] G. Andrew, R. Arora, J. Bilmes et al., "Deep canonical correlation analysis," in *Proceedings of the International Conference on Machine Learning*, pp. 1247–1255, Atlanta, GA, USA, June 2013.

[25] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3846–3853, New York City, NY, USA, July 2016.

[26] Y. Wei, Y. Zhao, C. Lu et al., "Cross-modal retrieval with CNN visual features: a new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, 2016.

[27] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, Bellevue, WA, USA, July 2013.

[28] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[29] A. Sharma, A. Kumar, H. Daume et al., "Generalized multiview analysis: a discriminative latent space," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2160–2167, IEEE, Providence, RI, USA, June 2012.

[30] J. Chi and Y. Peng, "Zero-shot cross-media embedding learning with dual adversarial distribution network," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 847–850, 2019.

[31] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2222–2230, Lake Tahoe, CA, USA, December 2012.

[32] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 405–420, 2018.

[33] X. Huang, Y. Peng, and M. Yuan, "MHTN: modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 1047–1059, 2018.

[34] J. Yu, Y. Lu, Z. Qin et al., "Modeling Text with Graph Convolutional Network for Cross-Modal Information retrieval," in *Proceedings of the Pacific Rim Conference on Multimedia*, pp. 223–234, Springer, Hefei, China, September 2018.

[35] X. Xu, A. Shimada, R. Taniguchi et al., "Coupled dictionary learning and feature mapping for cross-modal retrieval," in *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, Turin, Italy, July 2015.

[36] X. Xu, Y. Yang, A. Shimada, R.-I. Taniguchi, and L. He, "Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 847–850, ACM, Brisbane, Australia, October 2015.

[37] Z. Ye and Y. Peng, "Multi-scale correlation for sequential cross-modal hashing learning," in *Proceedings of the 2018 ACM multimedia conference on multimedia conference*, pp. 852–860, ACM, Seoul, Republic of Korea, March 2018.

[38] Y. Yang, F. Nie, D. Xu et al., "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, 2011.

[39] J. Yan, H. Zhang, J. Sun et al., "Joint graph regularization based modality-dependent cross-media retrieval," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 3009–3027, 2018.

[40] Y. Wei, Y. Zhao, Z. Zhu et al., "Modality-dependent cross-media retrieval," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 4, p. 57, 2016.

[41] J. Krapac, M. Allan, J. Verbeek et al., "Improving web image search results using query-relative classifiers," in *Proceedings of the 2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 1094–1101, IEEE, San Francisco, CA, USA, June 2010.

[42] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.

[43] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2010–2023, 2016.

[44] J. Wu, Z. Lin, and H. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 917–920, ACM, Tokyo, Japan, August 2017.

[45] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2018.

[46] F. Shang, H. Zhang, J. Sun, L. Liu, and H. Zeng, "A cross-media retrieval algorithm based on consistency preserving of collaborative representation," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 22, no. 2, pp. 280–289, 2018.

[47] Y. Qi, H. Zhang, B. Zhang et al., "Cross-media retrieval based on linear discriminant analysis," *Multimedia Tools and Applications*, vol. 78, pp. 1–20, 2018.

[48] Y. Wu, S. Wang, W. Zhang et al., "Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval," in *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 823–828, IEEE, Hong Kong, China, July 2017.