IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 15, 2019, accepted December 31, 2019, date of publication January 13, 2020, date of current version January 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2966220

Semantic Consistency Cross-Modal Retrieval With Semi-Supervised Graph Regularization

GONGWEN XU[1], XIAOMEI LI[2], AND ZHIJUN ZHANG[3]

[1]Business School, Shandong Jianzhu University, Jinan 250101, China
[2]The Second Hospital, Shandong University, Jinan 250013, China
[3]Computer Science and Technology School, Shandong Jianzhu University, Jinan 250101, China

Corresponding author: Gongwen Xu (xugongwen@163.com)

This work was supported in part by the Shandong Provincial Key Research and Development Program under Grant 2016GGX101035 and Grant 2019GGX101004, in part by the Shandong Education Department Teaching Reform Project under Grant Z2016M014, Grant Z2016M016, Grant Z2016Z013, and Grant M2018x197, and in part by the China Scholarship Council under Grant 201809995006.

ABSTRACT Most of the existing cross-modal retrieval methods make use of labeled data to learn projection matrices for different modal data. These methods usually learn the original semantic space to bridge the heterogeneous gap, ignoring the rich semantic information contained in unlabeled data. Accordingly, a semantic consistency cross-modal retrieval with semi-supervised graph regularization (SCCMR) algorithm is proposed, which integrates the prediction of labels and the optimization of projection matrices into a unified framework to ensure that the solution obtained is globally optimal. At the same time, the method uses graph embedding to consider the nearest neighbors in the potential subspace of paired images and texts as well as images and texts with the same semantics. l_{21}-norm constraint is applied to the projection matrices to select the discriminative features for different modal data. The results show that our method outperforms several advanced methods on four commonly used cross-modal retrieval datasets.

INDEX TERMS Cross-modal retrieval, semi-supervised, graph regularization, subspace learning.

I. INTRODUCTION

With the arrival of the big data era, data such as texts, images, audio, and videos have experienced an explosive growth on the Internet. Users need a variety of hybrid modalities of retrieval, so that retrieval methods based on single modal [1]–[3] data can no longer meet people's needs. The cross-modal retrieval [4]–[6] technology has emerged in a historic moment and, owing to its great significance in both theoretical research and practical applications, it will gradually become the most popular research direction in the field of information retrieval.

Cross-modal retrieval technology, as its name implies, is a technology that different modal data could retrieve each other. As shown in Figure 1, when we show a picture of a tiger, there will appear text describing the picture, as well as audio and video related to tigers. Cross-modal retrieval is popular among users because of its more comprehensive and detailed description of content. However, the following chanllenges are emerged: (1) the underlying features and dimensions of different modal data are different, and it is impossible to measure their similarity directly; how to bridge the heterogeneous gap between underlying features and high-level semantics is the most important issue in cross-modal retrieval; (2) as

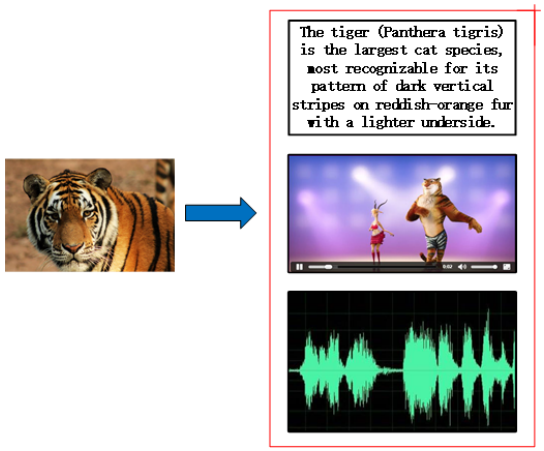


FIGURE 1. The description of cross-modal retrieval. When a picture of a tiger as a query instance is submitted, the text description of the tiger, audio of the tiger's call and the video of the tiger will be returned.

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

data of different modalities coexist, there are many types of noise and redundancies; how to select discriminant features and remove noise effectively is also a problem to be solved urgently in cross-modal retrieval; (3) in practice, multi-modal data are complex and diverse, but the labeled data account for only a small part; how to effectively utilize the rich semantic information contained in labeled and unlabeled data has become a major challenge in cross-modal retrieval.

To solve the first problem, many methods [7]–[9] learn a potential subspace for different modal data. In this subspace, multi-modal data have the same dimension representation characteristics, hence they can directly measure the similarity of different modal data. For example, a series of algorithms [10], [11] based on canonical correlation analysis (CCA) [12] projected different modal data into a shared subspace via subspace mapping. Zhang and Chen [13] proposed kernal CCA (KCCA) which mapped data of different modalities into high-dimensional space to learn the semantic concepts corresponding to images and text. However, these subspace learning methods focus on the learning of projection matrices and ignore the selection of discriminative features, which can't achieve satisfactory results. In the proposed method of this article, l_{21}-norm constraint is applied to the projection matrices to select the discriminative features for different modal data.

To solve the second problem, many methods [14]–[16] use different constraints, such as the l_1-norm [17], [18], to obtain sparse representations of different modalities, which can improve the speed of cross-modal retrieval. F-norm constraint [19], [20] is used to remove the noise of different modalities and obtain more discriminant features. Different from these above methods which concentrate on constraining the projection matrices, our method makes full use of the relationship of intra-modal data and inter-modal data, and can achieve better result.

To solve the third problem, the joint feature selection and subspace learning (JFSSL) method [21] made full use of the structural information of labeled and unlabeled data, and effectively improves the performance of cross-modal retrieval by constructing a series of graphs. Semi-supervised methods [22], [23] have also been developed, where the unlabeled data are tagged with pseudo-labels under the help of the labeled data, and the projection matrix is learned by using the original labels and pseudo-labels for different modal data. Compared with these above methods, our method can dynamically correct the prediction labels, and can obtain the global optimal result.

Deep neural network (DNN) [24] has drawn more and more attention because of its multi-layer nonlinear projection property. Therefore, DNN-based methods have become the focus of cross-modal retrieval research. In particular, Generative adversarial network (GAN) has widely used in cross-modal retrieval because of its powerful characteristics of modeling the underlying feature. Wang et al. [25] proposed ACMR (Adversarial Cross-Modal Retrieval), and the core idea of this method is to obtain an effective shared subspace through the confrontation mechanism between feature projector and modality classifier, so as to retrieval for different modalities in this subspace. Xu *et al.* [26] introduced the idea of metric learning into the process of adversarial learning. Not only the statistical characteristics of visual modality and textual modality are preserved, but also the correlations between different modalities are maximized. However, owing to the characteristic of high time complexity, the above methods are limited in promotion. In addition, they cannot effectively utilize unlabeled data.

Accordingly, in this paper we propose a semantic consistency cross-modal retrieval with semi-supervised graph regularization (SCCMR) algorithm (Figure 2), which can simultaneously take into account the heterogeneous gap,

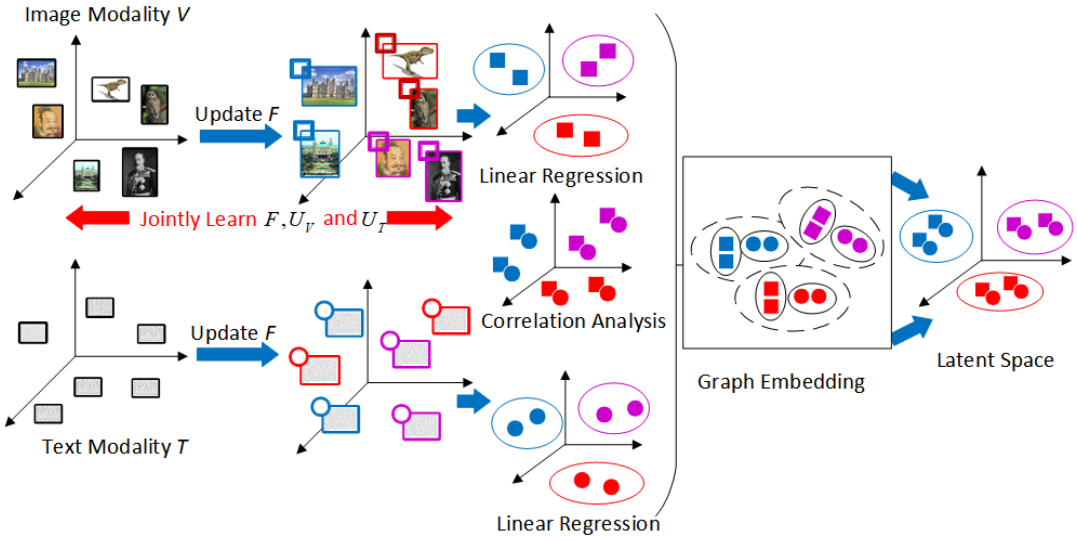


FIGURE 2. The flowchart of our proposed method.

the selection of discriminant information, and the semantic information of unlabeled data. We also propose an effective optimization algorithm. Experiments show that our algorithm outperforms several advanced cross-modal retrieval methods on four datasets. The main contributions of this paper are as follows:

1. A novel semi-supervised cross-modal retrieval algorithm is proposed, which can simultaneously optimize the prediction labels and projection matrices of different modalities and use the original semantic labels to correct the prediction labels, ensuring that the obtained prediction label matrix is globally optimal.

2. The graph embedding method ensures the approximation of paired images and texts in potential subspace as well as their approximation with the same semantics.

3. When learning a potential space, we use the l_{21}-norm to select features with strong correlation and discrimination. Further, we propose an effective iterative algorithm to ensure the convergence of the objective function.

The remaining chapters of this paper are organized as follows. In Section 2, we introduce the latest cross-modal retrieval methods. In Section 3, we introduce our algorithm and its optimization in detail. In Section 4, the experimental results are analyzed in detail on four datasets. And in Section 5, we summarize the main contributions of this paper.

II. RELATED WORKS

The key challenge of cross-modal retrieval technology [27] is to realize the correlation analysis of different modal data. In this process, different constraints are taken into account to improve the retrieval accuracy. Based on this, domestic and overseas scholars have developed so many cross-modal retrieval methods, such as subspace learning methods, dictionary learning methods, and pseudo-label construction methods.

Subspace learning methods are the earliest and most widely used in the cross-modal field. Research on these methods provides a foundation for other cross-modal retrieval methods. Traditional subspace learning methods involve projecting images and texts into the original semantic space. For example, Wei *et al.* [28] proposed a modality-dependent cross-media retrieval (MDCR) method, which is task-based and is used to learn different projection matrices for different cross-modal retrieval tasks. Because of the noise in the original semantic labels, many scholars are devoted to finding an orthogonal space to provide more accurate common representations for image and text learning. Wu *et al.* [29] used orthogonal constrained spectral regression to learn potential subspaces, which greatly improved the accuracy of cross-modal retrieval. However, this method only solves the problem of similarity measurement of different modal data and cannot select more discriminant data and effectively consider the internal structure of different modal data. Consequently, the results achieved so far cannot meet people's expectations.

Because the dimension of data of different modalities are so different, a lot of information will be lost if only seeking low-dimensional common subspaces for different modalities. Especially for high-dimensional image modalities, it is difficult to maintain the integrity of information. Dictionary learning methods can learn sparse coefficients for data of different modalities, which not only reduce the computational difficulty, but also improve the accuracy of cross-modal retrieval. Zhuang *et al.* [30] proposed a supervised coupled dictionary learning with group structures algorithm, which took advantage of dictionary to process different modal data. Meanwhile, the common structure was found by labeling information of similar data within modal data. Deng et al. [31] continued to explore dictionary learning and proposed a discriminative dictionary learning method, which uses common label alignment to learn different modalities of semantic mapping, thus, improving the retrieval performance. Xu *et al.* [17] learned the corresponding dictionary for different modal data, obtained its sparse representation, and then projected the sparse representations of different modal data into a common subspace for cross-modal retrieval. However, the above dictionary learning methods only use dictionaries to learn common representation for different modal data, and only use labeled data, ignoring the huge information contained in unlabeled data, preventing them from achieving better results.

Because unlabeled data contain vast information, many methods have been developed to construct pseudo-labels for unlabeled data of different modalities, and then learning projection matrices by using pseudo-labels and original semantic labels. Xu *et al.* [32] proposed a pseudo-label learning algorithm based on semantic consistency preservation. The main concept of the algorithm is to use labeled data to learn class centers, and then calculate pseudo-labels for different modalities. Finally, the original labels with labeled data and the pseudo-labels without labeled data are used as different modal learning projection matrices. Then, the common representation space is obtained for similarity measurement. A similar concept is used in the semi-supervised distance consistency preservation algorithm proposed by Dong *et al.* [33] Although both of the above algorithms use unlabeled data, they have the following disadvantages: (1) the unlabeled data is learned from the internal structure of labeled data and cannot mine the physical structure of unlabeled data effectively; (2) the pseudo-label is constructed for unlabeled data first, and then the projection matrix is produced. The pseudo-label and projection matrix cannot be optimized simultaneously. Therefore, the result is not globally optimal.

To overcome the shortcomings of the above methods, we propose the SCCMR method. Not only can the intrinsic structure relationship between labeled data and unlabeled data be fully excavated, but the prediction labels and projection matrices of different modalities can also be optimized to further improve the performance of cross-modal retrieval.

III. SEMANTIC CONSISTENCY CROSS-MODAL RETRIEVAL WITH SEMI-SUPERVISED GRAPH REGULARIZATION

In this section, we describe the SCCMR method and an effective optimization algorithm to optimize the objective

function. Detailed formal definitions and explanations will also be described.

A. THE FORMAL DEFINITION

Given a labeled data set $G_1 = \{(v_i, t_i, y_i)_{i=1}^{n_1}\}$, where $v_i \in R^{d_v}$ and $t_i \in R^{d_t}$ correspond to the original underlying features of the image and text, respectively, there are n_1 pairs of samples. (v_i, t_i) is an image-text pair with a common semantic $y_i \in R^c$, where c represents the number of categories of semantic concepts in the data set. We define a unlabeled data set $G_2 = \{(v_j, t_j)_{j=n_1+1}^{n}\}$, with $n - n_1$ visual and textual samples. Then, we define an image training set $V = [v_1, v_2, \ldots, v_{n_1}, v_{n_1+1}, \ldots, v_n] \in R^{d_v \times n}$ and a text training set $T = [t_1, t_2, \ldots, t_{n_1}, t_{n_1+1}, \ldots, t_n] \in R^{d_t \times n}$, d_v and d_t represent the dimensions of the underlying features of the image and text, respectively. $Y = [y_1, y_2, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_n] \in R^{c \times n}$ and $F = [f_1, f_2, \ldots, f_{n_1}, f_{n_1+1}, \ldots, f_n] \in R^{c \times n}$ represent a real label matrix and a prediction label matrix, respectively. In particular, $y_k = 0, \forall k = n_1 + 1, \ldots, n$.

B. PROPOSED METHOD

To make full use of the information of labeled and unlabeled data, we propose a semantic consistency cross-modal retrieval with semi-supervised graph regularization algorithm. This algorithm aims at learning different projection matrices for different modalities, in order to project the data of different modalities into a potential subspace for similarity measurements. Because unlabeled data lacks real labels, we cannot directly use the original semantic label matrix to achieve our goal. Therefore, we learn a predictive label matrix and use the real label matrix with labeled data to dynamically correct the predictive label matrix. To make predictive labels as close as possible to real labels, we define a graph embedding method:

$$\sum_{i=1}^{n} u_1 ||f_i - y_i||_F^2 = tr((F - Y)^T U (F - Y)) \tag{1}$$

where $U \in R^{c \times c}$ is a diagonal matrix, $U_{ii} = u_i$. And we define u_i as follows:

$$u_i = \begin{cases} +\infty, & labeled\ data \\ 0, & otherwise \end{cases} \tag{2}$$

Therefore, we can obtain the following final objective function:

$$\begin{aligned} \min_{U_V, U_T} \ & \alpha(||F - U_V^T V||_F^2 + ||F - U_T^T T||_F^2) \\ & + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} ||U_V^T V_i - U_T^T T_j||_2^2 \\ & + \beta(||U_V||_{21} + ||U_T||_{21}) \\ & + tr((F - Y)^T U (F - Y)) \end{aligned} \tag{3}$$

where α, λ and β are super parameters, which are used to control the weight of each item. $U_V \in R^{d_v \times c}$ and $U_T \in R^{d_t \times c}$ represent the projection matrices of images and texts, respectively. We use the l_{21}-norm to constrain the projection matrix. On the one hand, it can avoid data over-fitting, on the other hand, it can be used to select features with strong correlation and discrimination.

The adjacency matrix W_{ij} is defined as follows:

$$w_{ij} = \begin{cases} 1/N_t, & if\ v_i\ and\ t_j\ belong\ tothe\ same\ class\ t \\ 0, & otherwise \end{cases} \tag{4}$$

where N_t represents the number of t-th class. The first item of the objective function is a linear regression term, which guarantees that images and texts with the same semantics are approximated in potential subspaces. The second item is the correlation analysis item, which is used to ensure that image-text pairs are approximated in the potential subspace, and to further ensure the approximation of images and texts with the same semantics in the potential subspace. Traditional methods use the following paired constraints:

$$\lambda \sum_{i=1}^{N} \sum_{j=1}^{N} ||U_V^T V_i - U_T^T T_j||_2^2 \tag{5}$$

The above formula only guarantees the approximation of image-text pairs in potential subspaces and cannot achieve our goal.

C. OPTIMIZATION

We find that the objective function is convergent when one of the variables is updated while the others are fixed. Accordingly, we propose an effective iterative algorithm. First, we optimize $||U_V||_{21}$ and $||U_T||_{21}$. Following the procedure in [21], we take U_V as an example. We define $\varphi(x) = \sqrt{x^2 + \varepsilon}$, where ε is a smoothing term which can be expressed by a small constant to ensure the convergence of the iterative algorithm. Then, we can have $||U_V||_{21} = \sum_{b=1}^{d_i} \varphi(||u_V^b||_2)$ and the following formula can be obtained:

$$||U_V||_{21} = tr(U_V^T R_V U_V) \tag{6}$$

where $R_V = Diag(r_V)$. The b-th element of R_V is represented as: $r_V^b = 1/2||u_V^b||_2$, thus, r_V^b can be represented as follows:

$$r_V^b = \frac{1}{2\sqrt{|||u_V^b||_2^2 + \varepsilon}} \tag{7}$$

Similarly, $||U_T||_{21}$ is expressed as:

$$||U_T||_{21} = tr(U_T^T R_T U_T) \tag{8}$$

Then we fix F and U_T to update the value of U_V, we can obtain:

$$\begin{aligned} \min_{U_V} J(U_V) &= \alpha||F - U_V^T V||_F^2 + \beta||U_V||_{21} \\ &\quad + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} ||U_V^T V_i - U_T^T T_j||_2^2 \\ &= \alpha||F^T - V^T U_V||_F^2 + \beta||U_V||_{21} \\ &\quad + \lambda(tr(U_V^T V D V^T U_V) - tr(U_V^T V W T^T U_T) \\ &\quad - tr(U_T^T T W V^T U_V)) \end{aligned} \tag{9}$$

where D is the degree matrix corresponding to the adjacent matrix W.

The partial derivative of the above formula for U_V is expressed as follows:

$$\frac{\partial J(U_V)}{\partial U_V} = -2\alpha V(F^T - V^T U_V) + 2\beta R_V U_V + 2\lambda V(DV^T U_V - WT^T U_T) \quad (10)$$

We can get the following formula by making the derivative equal to zero.

$$U_V = (\alpha VV^T + \lambda VDV^T + \beta R_V)^{-1}(\alpha VF^T + \lambda VWT^T U_T) \quad (11)$$

Similarly, we derive the partial derivative of U_T and obtain the following formula:

$$\frac{\partial J(U_T)}{\partial U_T} = -2\alpha T(F^T - T^T U_T) + 2\beta R_T U_T + 2\lambda T(DT^T U_T - WV^T U_V) \quad (12)$$

Then,

$$U_T = (\alpha TT^T + \lambda TDT^T + \beta R_T)^{-1}(\alpha TF^T + \lambda TWV^T U_V) \quad (13)$$

Finally, we fix U_V, U_T, update F, and obtain the following formula:

$$\frac{\partial J(F)}{\partial F} = \alpha(F - U_V^T V + F - U_T^T T) + UF - UY = 0 \quad (14)$$

Thus,

$$F = (2\alpha I + U)^{-1}(\alpha(U_V^T V + U_T^T T) + UY) \quad (15)$$

The iterative algorithm for SCCMR is formulated in Algorithm 1.

Algorithm 1 Iterative Algorithm for SCCMR

Input: Image feature matrix $V = [v_1, v_2, \ldots, v_{n_1}, v_{n_1+1}, \ldots, v_n] \in R^{d_v \times n}$;
Text feature matrix $T = [t_1, t_2, \ldots, t_{n_1}, t_{n_1+1}, \ldots, t_n] \in R^{d_t \times n}$;
Real label matrix $Y = [y_1, y_2, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_n] \in R^{c \times n}$;
Parameters α, λ, β.

1: Initialize U_V, U_T , F as random matrices for iteration $k = 0$;
2: **repeat**
3: Updata U_V with U_T and F;
4: Updata U_T with U_V and F;
5: Updata F with U_V and U_T;
6: Set $k = k + 1$.
7: **until:** Objective function of Eq. (2) converges.

Output: Projection matrices $\boldsymbol{U}_V$, $\boldsymbol{U}_T$, predictive label matrix F.

IV. EXPERIMENTS

A. DATASETS

For the experiments, we used the following datasets:

Wikipedia dataset is the first open-access dataset in the field of cross-modal retrieval. It contains 2866 image-text pairs, each of which corresponds to one of ten semantics. We randomly selected 2173 pairs of data from the set to form the training set and 693 pairs of data to form the testing set. In this dataset, the text is represented by a 100-dimensional latent dirichlet allocation (LDA) feature and the image is represented by a 4096-dimensional convolutional neural network (CNN) feature.

Pascal Sentence dataset is a very popular dataset in the field of cross-modal retrieval. It contains 1000 pairs of image-text pairs from 20 semantic categories. Among them, 600 pairs are used for training and 400 pairs are used for testing. For text, we first obtain the 300-dimensional features of the text, and then the 100-dimensional features representation of the text based on LDA. For image, we use 4096-dimensional CNN visual features to represent them.

INRIA-Websearch dataset is a large-scale dataset in the field of cross-modal retrieval. It consists of 71478 image-text pairs, divided in 353 semantic categories. In particular, for this study, we selected 100 categories with the largest number of samples as the dataset. Among them, 10332 samples constitute the training set and 4366 samples constitute the testing set. In this dataset, the text is represented by 1000-dimensional LDA features. For image, 4096-dimensional CNN visual features are used.

NUS-WIDE-10k dataset is a very widely-used cross-modal dataset chosen from the 10 largest categories in NUS-WIDE dataset. Furthermore, 8000/2000 image-text pairs are utilized for training/testing. The images are represented by 4096-dimensional VGG vectors while the texts are represented by 1000-dimensional bag of words(BoW) vectors.

The statistical characteristics of the above four datasets are shown in Table 1.

B. EXPERIMENTAL SETTINGS AND COMPARED METHODS

1) EVALUATION METRICS

In this paper, we introduce two subtasks of cross-modal retrieval: image query text (I2T) and text query image (T2I). To further verify the effectiveness of our algorithm on the four datasets, we also performed the experiments with the three most common evaluation metrics: mean average precision (MAP), precision-recall curve (PR curve), and precision of each class. Detailed description of the proposed method is as follows:

We calculate the average precision value first:

$$AP = \frac{1}{R}\sum_{k=1}^{N} p(k)\delta(k) \quad (16)$$

where N is the number of returned samples, R represents the relevant data samples. $p(k)$ is the precision of the top k

TABLE 1. Key statistics of datasets.

Datasets	Wikipedia	Pascal Sentence	INRIA-Websearch	NUS-WIDE-10k
Database	2866	1000	14698	10000
Training	2173	600	10332	8000
Query	693	400	4366	2000
Visual feature	CNN(4096-D)	CNN(4096-D)	CNN(4096-D)	VGG(4096-D)
Text feature	LDA(100-D)	LDA(100-D)	LDA(1000-D)	BoW(1000-D)

retrieved samples. $\delta(k) = 1$ indicates that the k-th instance is consistent with the query term, otherwise $\delta(k) = 0$. Then, the MAP can be defined as:

$$MAP = \frac{1}{N}\sum_{i=1}^{N} AP(q_i) \tag{17}$$

where q_i denotes the i-th query instance.

2) PARAMETER TUNING

In our proposed objective function, there are three parameters, namely α, λ and β. Among them, α and λ control the weights of the linear regression and correlation analysis. Taking the image query text task on the Wikipedia dataset as an example, we use the grid search method to obtain the optimal parameters. As can be seen from Figure 3, when $\alpha = 0.05$, $\lambda = 0.05$, and $\beta = 0.1$, the MAP value is maximum. Using the same method, we can also obtain the optimal parameters on the other three datasets.

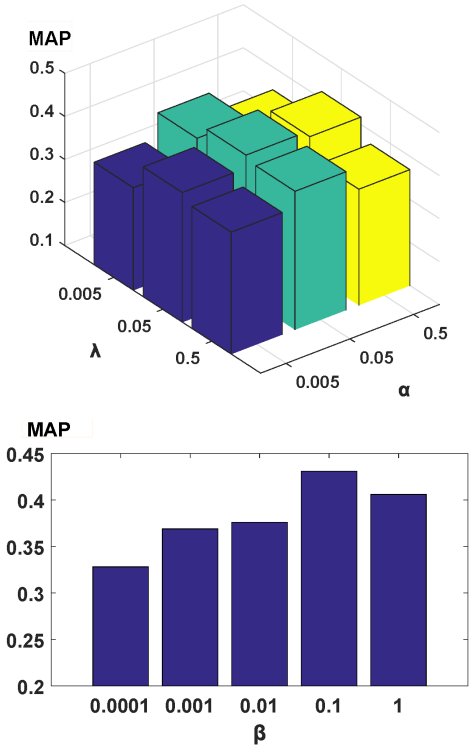


FIGURE 3. Experimental results of our method with different values of model parameters on Wikipedia dataset.

3) COMPARED METHODS

In the experiment, we compared our SCCMR method with the following twelve advanced cross-modal retrieval methods:

1. **Tree-view CCA (CCA-3V/T-V CCA)** [34] utilizes image view, text view, and semantic view. The semantic view can capture high-level semantic information effectively and help to obtain better image and text representation in the potential subspace.

2. **Generalized Multiview Linear Discriminant Analysis (GMLDA) and Generalized Multiview Marginal Fisher Analysis (GMMFA)** [35] are extensions of Linear Discriminant Analysis (LDA) [36] and Marginal Fisher Analysis (MFA) [37], respectively. They all use generalized multi-view analysis (GMA), GMLDA = GMA + LDA and GMMFA = GMA + MFA.

3. **JFSSL** chooses features with strong correlation and discrimination through l_{21}-norm, and learns a common subspace for images and texts. The nearest neighbor relationship between image and text in the subspace is maintained by regression.

4. **MDCR** is a modality-dependent cross-modal retrieval method, which assigns different projection matrices to different cross-modal retrieval tasks, and greatly improves the efficiency of cross-modal retrieval.

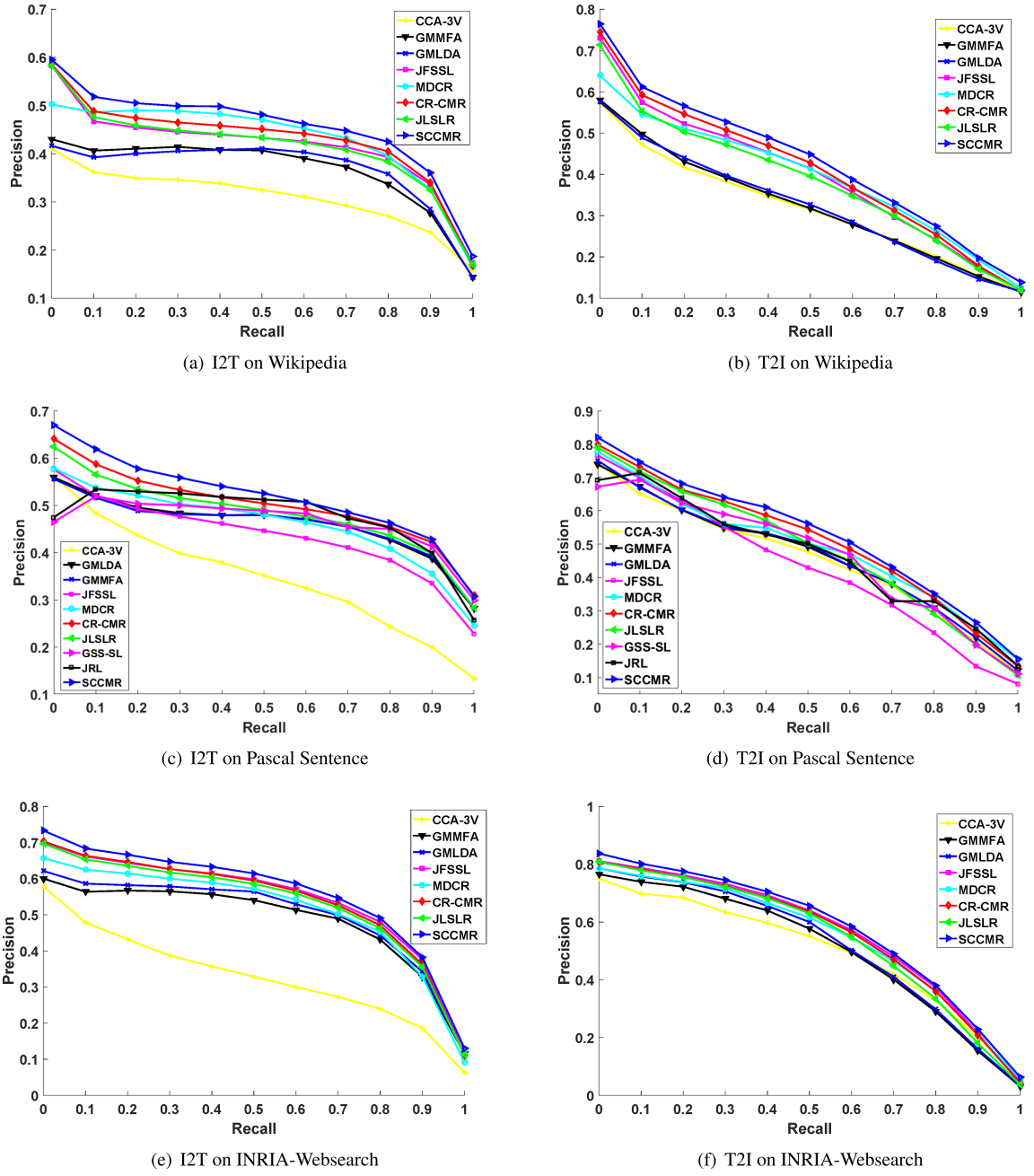
5. **Collaborative representation cross-media (CR-CMR)** [38] is an advanced cross-modal retrieval method for dictionary learning. It learns different dictionaries for images and texts, and obtains corresponding collaborative representations. At the same time, this method maintains semantic consistency.

6. **Joint latent subspace learning and regression (JLSLR)** [29] uses spectral regression when learning potential subspaces. Through orthogonal constraints, JLSLR learns more accurate orthogonal space than the original semantic space. In this space, maintaining the relevance of the image and text is also considered.

7. **GSS-SL** [14] and **JRL** [39] are two state-of-the-art semi-supervised methods in the field of cross-modal retrieval.

8. **Multimodal DBN** [40] takes advantage of Deep Boltzmann Machine to learn a generative model which contains diverse input modalities, and it can obtain the joint representations of different modalities.

9. **Bimodal-AE** [41] and **Corr-AE** [42] utilize Autoencoder to construct model. Compared with Bimodal-AE, Corr-AE gets a better performance by adding association constraints to the representation layer of the two single-modality



(a) I2T on Wikipedia

(b) T2I on Wikipedia

(c) I2T on Pascal Sentence

(d) T2I on Pascal Sentence

(e) I2T on INRIA-Websearch

(f) T2I on INRIA-Websearch

FIGURE 4. Precision-recall curves on three datasets.

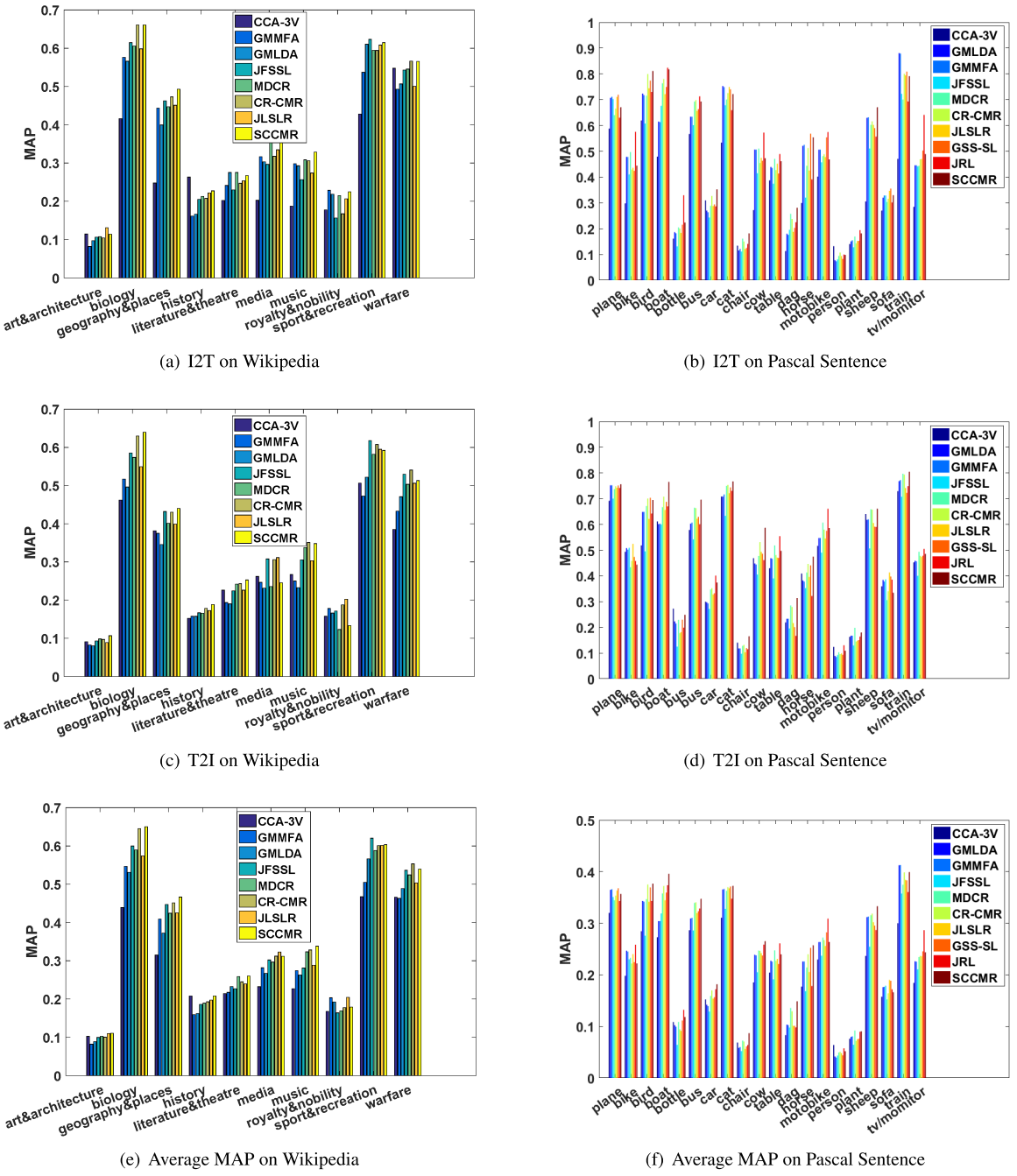
autoencoders, which pay more attention to the common parts of the two modalities when reconstruct the input data.

C. ANALYSIS OF EXPERIMENTAL RESULTS

In this section, we will describe the performance analysis of our method on four common datasets with the three evaluation methods mentioned above. Table 2 gives the MAP values of nine advanced cross-modal retrieval methods and our method on three datasets. Table 3 gives the MAP values of three DNN-based methods and a advanced semi-supervised method on NUS-WIDE-10k dataset. Figure 4 shows the PR value of two cross-modal retrieval subtasks of SCCMR on the Wikipedia, Pascal Sentence and INRIA-Websearch datasets. Figure 5 shows the MAP values for each class of SCCMR on the first two datasets. Detailed analyses are given below:

1) PERFORMANCE ON WIKIPEDIA DATASET

As can be seen from Table 2, our method achieves higher MAP scores on the Wikipedia datasets than the first eight methods. For the image query text task, our method is 0.7% higher than GSS-SL. For text query image task, our method is also 0.8% higher than CR-CMR method. However, there



(a) I2T on Wikipedia

(b) I2T on Pascal Sentence

(c) T2I on Wikipedia

(d) T2I on Pascal Sentence

(e) Average MAP on Wikipedia

(f) Average MAP on Pascal Sentence

FIGURE 5. **MAP scores for each class on two datasets.**

still exist a exception. Our SCCMR achieves a slightly lower accuracy than JRL. The reason is that JRL can benefit from both modality correlation and semantic information. In addition, it has powerful ability to resist noise. The reasons why our method is superior to the other methods are given below.

1. The CCA-3V, GMMFA, GMLDA, MDCR, and JLSLR only use labeled data, ignoring the rich information contained in the unlabeled data. The JFSSL uses unlabeled data only to construct graphs and does not fully mine their semantic information. Our method not only utilizes the structural information of unlabeled data to construct Laplacian graphs, but also fully extracts the rich semantic information in the unlabeled data, which is helpful to learn a superior potential subspace, significantly improving the cross-modal retrieval.

2. Although GSS-SL utilizes the unlabeled data to increase the diversity of training set, it cannot dynamically update prediction labels. However, our method not only ensures that the image-text pairs are approximate in the potential subspace, but also ensures the approximation of images and texts with the same semantics in the potential subspace. In particular,

TABLE 2. The MAP scores on Wikipedia dataset, Pascal Sentence dataset and INRIA-Websearch dataset. The best result in each column is marked with bold.

Methods	Wikipedia			Pascal Sentence			INRIA-Websearch		
	I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
CCA-3V	0.310	0.316	0.313	0.337	0.439	0.388	0.329	0.500	0.415
GMMFA	0.371	0.322	0.346	0.455	0.447	0.451	0.492	0.510	0.501
GMLDA	0.372	0.322	0.347	0.456	0.448	0.452	0.505	0.522	0.514
JFSSL	0.392	0.381	0.387	0.406	0.401	0.404	0.532	0.559	0.547
MDCR	0.419	0.382	0.401	0.449	0.475	0.462	0.520	0.551	0.535
CR-CMR	0.408	0.395	0.402	0.471	0.480	0.476	0.532	0.555	0.544
JLSLR	0.393	0.369	0.381	0.454	0.455	0.455	0.525	0.544	0.535
GSS-SL	0.424	0.384	0.404	0.468	0.464	0.466	0.531	0.557	0.544
JRL	**0.443**	**0.405**	**0.424**	0.479	0.462	0.471	0.526	0.543	0.535
SCCMR	0.431	0.403	0.417	**0.489**	**0.496**	**0.493**	**0.541**	**0.568**	**0.555**

TABLE 3. The MAP scores on NUS-WIDE-10k dataset. The best result in each column is marked with bold.

Methods	NUS-WIDE-10k		
	I2T	T2I	Avg
Multimodal DBN	0.201	0.259	0.230
Bimodal-AE	0.327	0.369	0.348
Corr-AE	0.366	**0.417**	0.392
JRL	0.426	0.376	0.401
SCCMR	**0.434**	0.386	**0.410**

it can utilize the original semantic labels to correct the prediction labels, which effectively improves the completeness of the model.

Figure 4(a) and (b) and Figure 5(a), (c), and (e) show the PR value of our method and the MAP values of each class. We can see that our method is still superior, which further verifies the effectiveness of our method.

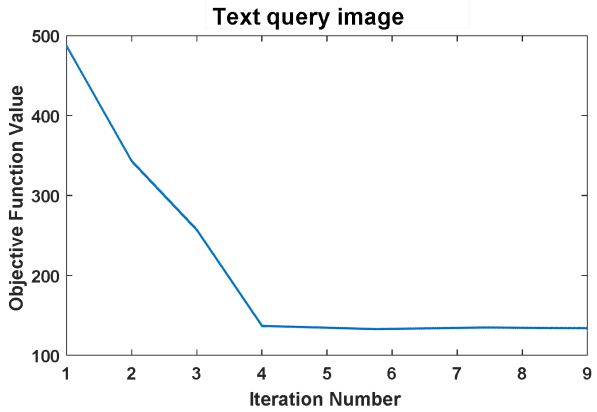
2) PERFORMANCE ON PASCAL SENTENCE DATASET AND INRIA-WEBSEARCH DATASET

As can be seen from Table 2, the average MAP scores of our method on Pascal Sentence datsaset and INRIA-Websearch dataset are 1.7% and 1.1% higher than those of the sub-optimal CR-CMR, respectively. Further, we can see that the MAP values of all methods on INRIA-Websearch dataset are higher than those on the Pascal Sentence dataset, which shows that the larger is the feature dimension, i.e., the larger the samples size, the better the performance will be. As can be seen from Figure 4, the PR value of our method on above two datasets are also significantly higher than those of other methods. In Figure. 5(b), (d), and (f) we can see that our method achieves the highest MAP score in most semantic categories. The above analyses further verify the effectiveness of our method.

3) PERFORMANCE ON NUS-WIDE-10K DATASET

In order to further verify the progressiveness and effectiveness of our SCCMR on NUS-WIDE-10k dataset, we compare our SCCMR with four methods with high accuracy, including three deep neural network methods Multimodal DBN, Bimodal-AE and Corr-AE. As shown in Table 3, benefiting from the graph embedding and l_{21}-norm constraints, SCCMR keeps obvious advantages with 4 compared methods on this dataset.

(a) I2T on Wikipedia



(b) T2I on Wikipedia

FIGURE 6. Convergence curves of the objective function values.

D. TIME-CONSUMING AND CONVERGENCE

As shown in Table 4, we can obtain that running time increases with the number of samples in the dataset. All experiments are implemented on Intel(R) Core(TM) i7-6700 CPU 3.40 GHz×2 machine with 24GB RAM. To test the convergence of our proposed iterative algorithm, we take

TABLE 4. The time-consuming of SCCMR on four datasets.

Dataset	Running time(s)
Wikipedia	168.31
Pascal Sentence	78.33
INRIA-Websearch	2836.44
NUS-WIDE-10k	308.16

the Wikipedia dataset as an example, and draw the line diagrams of image query text and text query image. As shown in Figure 6, we can see that our algorithm converges about the fourth iteration.

E. RETRIEVAL CASE ANALYSIS

Figure 7 shows examples of image query text and text retrieving image on Wikipedia dataset. Because it is difficult to see the semantics of text description, for better observation, we use the images corresponding to the texts to replace the text results in image query text. We can see that the first five values we retrieved are correct in both image query text and text query image subtask, which further verifies the effectiveness of our method.

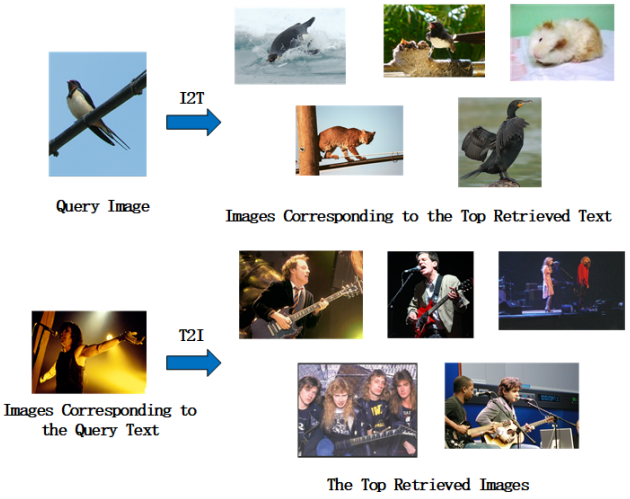


FIGURE 7. Two examples of image query text and text query image on Wikipedia dataset. For the example of image query text, we use the corresponding images of retrieved texts to demonstrate the results.

F. CONCLUSION

In this paper, the SCCMR method and an effective optimization algorithm used to optimize our objective function are proposed. The SCCMR makes full use of labeled and unlabeled data of different modalities, simultaneously optimizes the prediction labels and projection matrices of different modalities, and corrects the prediction labels with the original semantic labels to ensure that the obtained prediction label matrix is globally optimal. At the same time, the proposed method would choose the features with strong discrimination. The experimental results show that the method performs well on four commonly used cross-modal retrieval datasets.

REFERENCES

[1] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, Jul. 2018.

[2] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.

[3] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2868–2881, Jun. 2017.

[4] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.

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

[6] L. Wang, W. Sun, Z. Zhao, and F. Su, "Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval," *Signal Process.*, vol. 131, pp. 249–260, Feb. 2017.

[7] X. Zhao, C. Zhang, and Z. Zhang, "Distributed cross-media multiple binary subspace learning," *Int. J. Multimedia Inf. Retr.*, vol. 4, no. 2, pp. 153–164, Jun. 2015.

[8] K. Li, G.-J. Qi, J. Ye, and K. A. Hua, "Linear subspace ranking hashing for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1825–1838, Sep. 2017.

[9] J. Wu, Z. Lin, and H. Zha, "Joint dictionary learning and semantic constrained latent subspace projection for cross-modal retrieval," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2018, pp. 1663–1666.

[10] J. Shao, L. Wang, Z. Zhao, F. Su, and A. Cai, "Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval," *Neurocomputing*, vol. 214, pp. 618–628, Nov. 2016.

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

[12] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[13] H. Zhang and L. Chen, "Learning optimal data representation for cross-media retrieval," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1925–1928.

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

[15] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "PL-ranking: A novel ranking method for cross-modal retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 1355–1364.

[16] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.

[17] X. Xu, A. Shimada, R.-I. Taniguchi, and L. He, "Coupled dictionary learning and feature mapping for cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun. 2015, pp. 1–6.

[18] Y. Xu, Y. Sun, Y. Quan, and B. Zheng, "Discriminative structured dictionary learning with hierarchical group sparsity," *Comput. Vis. Image Understand.*, vol. 136, pp. 59–68, Jul. 2015.

[19] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1220–1233, Jun. 2017.

[20] M. Zhang, H. Zhang, J. Li, L. Wang, Y. Fang, and J. Sun, "Supervised graph regularization based cross media retrieval with intra and inter-class correlation," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 1–11, Jan. 2019.

[21] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[22] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. T. Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015.

[23] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[24] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 1–24, Feb. 2019.
[25] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2017.
[26] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.
[27] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2019.
[28] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, and S. Yan, "Modality-dependent cross-media retrieval," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 1–13, Mar. 2016.
[29] J. Wu, Z. Lin, and H. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 917–920.
[30] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2013.
[31] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 208–218, Feb. 2016.
[32] G. Xu, Z. Sang, and Z. Zhang, "Cross-media retrieval based on pseudo-label learning and semantic consistency algorithm," *Int. J. Performab. Eng.*, vol. 14, no. 9, p. 2219, 2018.
[33] X. Dong, E. Yu, M. Gao, L. Zhu, J. Sun, and H. Zhang, "Semi-supervised Distance Consistent Cross-modal Retrieval," in *Proc. Workshop Vis. Anal. Smart Connected Commun. (VSCC)*, 2017, pp. 25–31.
[34] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.
[35] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multi-view analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.
[36] A. Sharma and K. K. Paliwal, "A deterministic approach to regularized linear discriminant analysis," *Neurocomputing*, vol. 151, pp. 207–214, Mar. 2015.
[37] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal Fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2811–2821, Nov. 2007.
[38] F. Shang, H. Zhang, J. Sun, L. Liu, and H. Zeng, "A cross-media retrieval algorithm based on consistency preserving of collaborative representation," *J. Adv. Comput. Intell. Intell. Inform.*, vol. 22, no. 2, pp. 280–289, Mar. 2018.
[39] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.
[40] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, vol. 79, 2012.
[41] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.
[42] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 7–16.

GONGWEN XU received the B.Sc. degree in computer science from the Nanjing University of Posts and Telecommunications, in 1998, and the master's degree from Shandong University, in 2005.

He is currently an Associate Professor with the Business School, Shandong Jianzhu University. He has done a lot of work on cross-modal retrieval and machine learning during his Ph.D. program. He has chaired and participated in some research projects. He has published several research articles in machine learning field. His current research interests include cross-modal retrieval, image processing, and machine learning.

Dr. Xu is a Senior Member of CCF and a member of Shandong Association for Artificial Intelligence. He is also the Director of the Jinan Computer Federation.

XIAOMEI LI received the degree and the Ph.D. degree in clinical oncology from Shandong University, in 2004 and 2013, respectively. She is currently a Doctor with The Second Hospital, Shandong University. Her research interests are targeted therapy of cancer, intelligent analysis of medical images, and machine learning.

ZHIJUN ZHANG received the Ph.D. degree in artificial intelligence from Shandong Normal University, in 2016. He is currently a Professor with the Computer Science and Technology School, Shandong Jianzhu University. His research interests are artificial intelligence, image processing, and recommending systems.