Research Article

Semi-Supervised Cross-Modal Retrieval Based on Discriminative Comapping

Li Liu, Xiao Dong, and Tianshi Wang

School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

Correspondence should be addressed to Li Liu; liuli_790209@163.com

Received 8 May 2020; Revised 7 June 2020; Accepted 13 June 2020; Published 18 July 2020

Academic Editor: Jianquan Lu

Copyright © 2020 Li Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Most cross-modal retrieval methods based on subspace learning just focus on learning the projection matrices that map different modalities to a common subspace and pay less attention to the retrieval task specificity and class information. To address the two limitations and make full use of unlabelled data, we propose a novel semi-supervised method for cross-modal retrieval named modal-related retrieval based on discriminative comapping (MRRDC). The projection matrices are obtained to map multimodal data into a common subspace for different tasks. In the process of projection matrix learning, a linear discriminant constraint is introduced to preserve the original class information in different modal spaces. An iterative optimization algorithm based on label propagation is presented to solve the proposed joint learning formulations. The experimental results on several datasets demonstrate the superiority of our method compared with state-of-the-art subspace methods.

1. Introduction

In real applications, data are often represented in different ways or obtained from various domains. As a consequence, the data with the same semantic may exist in different modalities or exhibit heterogeneous properties. With the rapid growth of multimodal data, there is an urgent need for effectively analyzing the data obtained from different modalities [1–5]. Although there is much attention to the multimodal analysis, the most common method is to ensemble the multimodal data to improve the performance [6–9]. Cross-modal retrieval is an efficient way to achieve data from different modal data. The typical example is to take the image as a query to retrieve related texts (I2T) or to search images by utilizing the textual description (T2I). Figure 1 shows the detailed process for I2T and T2I tasks. The results obtained by cross-modal retrieval are more comprehensive compared with the results of traditional single-modality.

Generally, semantic gap and relevant measure impede the development of cross-modal retrieval. Although there are many approaches to solve this problem, the performance of these approaches still cannot achieve a satisfactory level. Therefore, the methods [10–16] are proposed to learn a common subspace by minimizing the pairwise differences to make different modalities comparable. However, task specificity and class information are often ignored, which leads to low-level retrieval performance.

To solve these problems mentioned above, this paper proposes a novel semi-supervised joint learning framework for cross-modal retrieval by integrating the common subspace learning, task-related learning, and class discriminative learning. Firstly, inspired by canonical correlation analysis (CCA) [7] and linear least squares, a couple of projection matrices are learnt by coupled linear regression to map original multimodal data to the common subspace. At the same time, linear discriminant analysis (LDA) and task-related learning (TRL) are used to keep the data structure in different modalities and the semantic relationship in the projection space. Furthermore, to mine the category information of unlabelled data, a semi-supervised strategy is utilized to propagate the semantic information from labelled data to unlabelled data. Experimental results on three public datasets show that the proposed method outperforms the previous state-of-the-art subspace approaches.
2. Related Work

Because cross-modal retrieval plays an important role in various applications, many subspace-based methods have been proposed by establishing the intermodal and intramodal correlation. Rasiwasia et al. [7] investigated the retrieval performance of various combinations of image features and textual representations, which cover all possibilities in terms of the two guiding hypotheses. Later, partial least squares (PLS) [17] has also been used for the cross-modal matching problem. Sharma and Jacobs [18] used PLS to linearly map images from different views into a common linear subspace, where the images have a high correlation. Chen et al. [19] solved the problem of cross-modal document retrieval by using PLS to transform image features into the text space, and the method easily achieved the similarity measure between two modalities. In [20, 21], the bilinear model and generalized multiview analysis (GMA) have been proposed and performed well in the field of cross-modal retrieval.

In addition to CCA, PLS, and GMA, Mahadevan et al. [22] proposed a manifold learning algorithm that can simultaneously reduce the dimension of data from different modalities. Mao et al. [23] introduced a cross-media retrieval method named parallel field alignment retrieval, which integrates a manifold alignment framework from the perspective of vector fields. Lin and Tang [24] proposed a common discriminant feature extraction (CDFE) method to learn the difference within each scattering matrix and between scattering matrices. Sharma et al. [21] improved LDA and marginal Fisher analysis (MFA) to generalized multiview LDA (GMLDA) and generalized multiview MFA (GMMFA) by extending from single-modality to multimodalities. Inspired by the semantic information, Gong et al. [25] proposed a three-view CCA to deeply explore the correlation between features and their corresponding semantics in different modalities.

Furthermore, other methods, such as dictionary learning, graph-based learning, and multiview embedding, are proposed for the cross-modal problem [26–29]. Zhuang et al. [30] proposed SiM2 by adding a group sparse representation to the pairwise relation learning to project different modalities into a common space. Xu et al. [31] proposed that dictionary learning and feature learning should be combined to learn the projection matrix adaptively. Deng et al. [32] proposed a discriminative dictionary learning method with the common label alignment by learning the coefficients of different modalities. Wei et al. [33] proposed a modal-related method named MDCR to solve the modal semantic problem. Wu et al. [34] utilized spectral regression and a graph model to jointly learn the minimum error regression and latent space. Wang et al. [35] proposed an adversarial learning framework, which can learn modality-invariant and discriminative representations of different modalities. And in this framework, the modality classifier and the feature projector compete with each other to obtain a better pair of feature representations. Cao et al. [36] used multiview embedding to obtain latent representations for visual object recognition and cross-modal retrieval. Zhang et al. [37] utilized a graph model to learn a common space for cross-modal by adding the relationship of intraclass and interclass in the projection process.

The main purpose of these methods is to solve the correlation of distance measure, but the class information and task specificity are not well solved. Therefore, how to solve the two problems at the same time for different tasks is particularly important. Based on the idea, we learn two couples of projections for different retrieval tasks and apply a linear discriminative constraint to the projection matrices. To achieve this goal, we combine task-related learning with linear discriminative analysis through semi-supervised label propagation. Figure 2 shows the flowchart of our method. Experimental results on three open cross-modal datasets demonstrate that our cross-modal retrieval method outperforms the latest methods.

3. Methodology

To improve the retrieval performance, we introduce the discriminative comapping and pay more attention to different retrieval tasks and class information preservation. Here, we focus on the retrieval of I2T and I2T, and it is easy to expand our method to the retrieval of other modalities.
3.1. The Objective Function. Define image data as $I = \{I_i; I_u\} \in \mathbb{R}^{n \times p}$ and text data as $T = \{T_i; T_u\} \in \mathbb{R}^{n \times q}$ separately, where $I_i \in \mathbb{R}^{n \times p}$ and $T_i \in \mathbb{R}^{n \times q}$ denote the labelled image $I_i$ and its text with $p$ dimensions, and $I_u \in \mathbb{R}^{n \times p}$ and $T_u \in \mathbb{R}^{n \times q}$ represent the unlabelled image $I_u$ and its text with $q$ dimensions. Let $D = \{I_i, T_i\}_{i=1}^n$ be $n$ pairs of image and text documents, where $D_l = \{I_i, T_i\}_{i=1}^n$ and $D_u = \{I_u, T_u\}_{i=1}^n$ denote the labelled and unlabelled documents, respectively. $S = [S_l; S_u] \in \mathbb{R}^{c \times d}$ is the semantic matrix where $c$ is the category number, $S_i$ is the label of labelled data with one-hot coding, and $S_u$ is the pseudo-label of unlabelled data. The goal of our method is to learn two couples of projection matrices that project data from different modalities into a common space for different tasks. Then, the cross-modal retrieval can be performed in the common space.

We propose a novel modal-related projection strategy based on semi-supervised learning for task specificity. Here, the pairwise closeness of multimodal data and the semantic projection are combined into a unified formulation. For I2T and T2I, the minimization forms are obtained as follows:

$$f_1(V, W) = \min_{V, W} \| IV^T - TW^T \|^2_F + \| IV^T - S \|^2_F,$$  \hspace{0.5cm} (1)

$$f_2(V, W) = \min_{V, W} \| IV^T - TW^T \|^2_F + \| TW^T - S \|^2_F,$$  \hspace{0.5cm} (2)

where $V$ and $W$ stand for the projection matrices for modalities $I$ and $T$, respectively.

The linear discriminant constraint to equations (1) and (2) is introduced to preserve the class information in the latent projection subspace. We denote $m$ as the mean of the labelled samples in the $i$th class and $\mu$ as the mean of all labelled samples. The intraclass scatter matrix can be defined as $S_w = \sum_{i=1}^{c} \left( P(i)/n_i \right) \sum_{x_k \in \text{class}_i} (x_k - m_i)(x_k - m_i)^T$, and the total scatter matrix can be represented as $S = \sum_{i=1}^{c} P(i)(\bar{m}_i - m)(\bar{m}_i - m)^T$. The objective function is represented as follows:

$$\min_{W \in \mathbb{R}^{d \times k}} \frac{\text{tr}(W^T S_w W)}{\text{tr}(W^T S W)}$$  \hspace{0.5cm} (3)

where $W \in \mathbb{R}^{d \times k}$ is the projection matrix and $d$ is the dimension of the basic vector.

According to equation (3), the linear discriminant constraint can be transformed into $WS_w W^T$, where $S_w$ is the intra-class scatter of $I$ is represented as $S_{I, I}$ and the inter-class scatter of $I$ as $S_I$. Under the multimodal condition, our method utilizes LDA projections to preserve class information of each modal. The corresponding formula is as follows:

$$\min_{V, W} V^T A V + W^T B W,$$  \hspace{0.5cm} (4)

where $A$ and $B$ denote $S_{I, I} - \gamma I, I$ and $S_{I, I} - \gamma I, I$ separately.

We add equation (4) to equations (1) and (2), respectively, and then get the objective functions of I2T and T2I in the following:

$$F_1(V, W) = \min_{V, W} \frac{\| IV^T - TW^T \|^2_F + (1 - \lambda)\| IV^T - S \|^2_F}{\text{tr}(W^T B W)},$$  \hspace{0.5cm} (5)

$$F_2(V, W) = \min_{V, W} \frac{\| IV^T - TW^T \|^2_F + (1 - \lambda)\| TW^T - S \|^2_F}{\text{tr}(W^T B W)},$$  \hspace{0.5cm} (6)
where \( \lambda \) is a tradeoff coefficient to balance pairwise information and semantic information and \( \mu_1 \) and \( \mu_2 \) are regularization parameters to balance the structure information of the image and text. According to equations (1) and (2), the structure projection of \( I \) and \( T \) is the same as the semantic projection. Consequently, our method can bridge the feature and semantic spaces. This can decrease the loss of projection and improve the performance of cross-modal retrieval.

We introduce the semi-supervised learning strategy. To propagate the label information from the labelled data, we utilize the radial basis function (RBF) kernel to evaluate the pairwise similarities between the unlabelled data after projection, and then the similarities are regarded as the label information to be updated in the optimization process until the results converge. For any data \( x_i \) and \( x_j \), the kernel function is defined as follows:

\[
k(x_i, x_j) = \exp \left( -\frac{\|x_i - x_j\|^2}{2\beta^2} \right),
\]

where \( \beta \) is the kernel parameter.

### 3.2. Algorithm Optimization

The objective functions of equations (5) and (6) are nonconvex, so the iteration method is used to update each variant when other variants are fixed alternatively.

For any matrix \( M \in \mathbb{R}^{N \times d} \), the partial derivative of equation (5) is represented as follows:

\[
\frac{\partial F_1}{\partial V} = 2\bigg( \lambda \big( V^T I - W T^T I \big) + (1 - \lambda) \big( V^T I - S^T I \big) \bigg) + \mu_1 \big( VA + V A^T \big),
\]

\[
\frac{\partial F_1}{\partial W} = 2\bigg( \lambda \big( W T^T T - V I^T T \big) + (1 - \lambda) \big( W T^T T - S^T T \big) \bigg) + \mu_2 \big( WB + W B^T \big).
\]

Similarly, the partial derivative of equation (6) is given as follows:

\[
\frac{\partial F_2}{\partial W} = 2\bigg( \lambda \big( W T^T T - V I^T T \big) + (1 - \lambda) \big( W T^T T - S^T T \big) \bigg) + \mu_2 \big( WB + W B^T \big),
\]

\[
\frac{\partial F_2}{\partial V} = 2\bigg( \lambda \big( V^T I - W T^T I \big) + (1 - \lambda) \big( V^T I - S^T I \big) \bigg) + \mu_1 \big( VA + V A^T \big).
\]

According to equations (8)–(11), our method can be solved by gradient descent. Algorithm 1 describes the optimization of cross-modal learning. After the projection matrices for the I2T and T2I tasks are obtained, \( I \) and \( T \) can be mapped to the common space where cross-modal retrieval is achieved.

### 4. Experiments

To evaluate the performance of the proposed method (MRRDC), we do comparison experiments with several other methods on three public datasets.

#### 4.1. Datasets

**4.1.1. Wikipedia Dataset.** This dataset consists of 2,866 image-text pairs labelled with one of 10 semantic classes. In this dataset, 2,173 pairs of data are selected as the training set, and the rest are the testing set. In our experiments, we use the public dataset [7] provided by Rasiwasia et al. (wiki-R), where images are represented by 128-dimensional SIFT description histograms [38], and the representation of the texts with 10 dimensions is derived from an LDA model [39]. At the same time, we also use the dataset provided by Wei et al. (wiki-W) [40], where 4,096-dimensional CNN features [41] are used to present images and 100-dimensional LDA features are utilized to denote the texts.

**4.1.2. Pascal Sentence Dataset [40].** This dataset consists of 1,000 image-text pairs with 20 categories. We randomly choose 30 pairs from each category as training samples and the rest as test samples. The image features are 4,096-dimensional CNN features, and the text features are 100-dimensional LDA features.

**4.1.3. INRIA-Websearch [42].** This dataset contains 71,478 pairs of image and text annotations from 353 classes. We remove some pairs which are marked as irrelevant and select the pairs that belong to any one of the 100 largest categories. Then, we get a subset of 14,698 pairs for evaluation. We randomly select 70% of pairs from each category as the training set (10,332 pairs), and the rest are treated as the testing set (4,366 pairs). Similarly, images are represented with 4,096-dimensional CNN features, and the textual tags are represented with 100-dimensional LDA features.

#### 4.2. Evaluation Metrics

To evaluate the performance of the proposed method, two typical cross-modal retrieval tasks are conducted: I2T and T2I. In the test phase, the projection matrices are used to map the multimodal data into the common subspace. Then, the data of different modalities can be retrieved. In all experiments, the cosine distance is adopted to measure the feature similarities. Given a query, the aim of each cross-modal task is to find the top-\( k \) nearest neighbors from the retrieval results.

The performance of the algorithms is evaluated by mean average precision (mAP), which is one of the standard information retrieval metrics. To obtain mAP, average precision (AP) is calculated by

\[
AP = \frac{1}{R} \sum_{i=1}^{T} P(i) \sigma(i),
\]

where \( R \) is the number of correlation data in the test dataset, \( P(i) \) is the precision of top \( r \) retrieval data, and if \( \sigma(i) = 1 \), the top \( r \) retrieval data are relevant; otherwise, \( \sigma(i) = 0 \). Then, the value of mAP can be obtained by averaging AP for all queries. The larger the mAP, the better the retrieval performance. Besides the mAP, the precision-recall curves and mAP performance for each class are used to evaluate the effectiveness of different methods.
4.3. Comparison Methods. To verify that our method has good performance, we compare our method with seven state-of-the-art methods, such as PLS [18], CCA [7], SM [7], SCM [7], GMLDA [21], GMMFA [21], MDCR [33], JLSLR [34], ACMR [35], and SGRCR [37].

PLS, CCA, SM, and SCM are typical methods that utilize pairwise information to learn a common latent subspace, where the similarity between different multimodal can be measured by metric methods directly. These kinds of approaches make the pairwise data in the multimodal dataset closer in the learned common subspace. GMLDA, GMMFA, and MDCR are based on the semantic category information via supervised learning, Due to the use of label information, these methods can easily learn a more discriminative subspace.

4.4. Experimental Setup. The parameters of the proposed MRRDC in Algorithm 1 for the retrieval tasks of I2T and T2I are set as follows: \( \lambda = 0.5 \), \( \gamma_1 = 0.3 \), \( \gamma_2 = 0.3 \), \( \mu_1 = 0.5 \), \( \mu_2 = 0.5 \), \( \gamma_1 = 0.5 \), \( \gamma_2 = 0.5 \), \( \epsilon_1 = 10^{-5} \), and \( \epsilon_2 = 0.001 \) on Wikipedia provided by Rasiwasia and INRIA-Websearch. On Wikipedia provided by Wei and Pascual, \( \lambda = 0.3 \), and the rest are the same with the above. In our experiment, learning rate \( \sigma \) is set \( 10^{-4} \).

4.5. Results and Analysis. Table 1 shows all the mAP scores achieved by PLS, CCA, SM, SCM, GMMFA, GMLDA, MDCR, and our method on wiki-R, wiki-W, Pascal Sentence, and INRIA-Websearch. We observe that our method outperforms its counterparts. This may be because the projection matrices preserve more discriminative class information via semi-supervised learning. The common subspace of our method is more discriminative and effective by further exploiting the class semantic of intramodality and intermodality similarity simultaneously. From Table 1, we also find that, in most cases, GMMFA, GMLDA, MDCR, and MRRDC always perform better than PLS, CCA, SM, and SCM, and images with CNN features have superiority compared with the shallow features. For the first result, this is because PLS, CCA, SM, and SCM only use pairwise information, but the other approaches add class information to their objective functions, which provides better separation between different categories in the latent common subspace. For the second result, this is due to the powerful semantic representation of CNN.

The precision-recall curves on wiki-R, wiki-W, Pascal Sentence, and INRIA-Websearch are plotted in Figure 3. Figure 4 shows the mAP scores of comparison approaches and our method, and the rightmost bar of each figure shows the average mAP scores. For most categories, the mAP of our method outperforms that of comparison methods. From these experimental results, we can draw the following conclusions:

(1) Compared with the current state-of-the-art methods, our method improves the average mAP greatly. Our method consistently outperforms compared methods, which is due to the factor that MRRDC learns projection matrices in task-related and linear discrimination ways for different modalities, where different modalities can preserve semantic and original class information. Besides, both labelled data
### Table 1: Retrieval performance (mAP) on three public datasets.

| Methods | I2T (%) | T2I (%) | Average (%) | I2T (%) | T2I (%) | Average (%) |
|---------|---------|---------|------------|---------|---------|------------|
| PLS     | 23.75   | 17.23   | 20.49      | 35.95   | 35.10   | 35.53      |
| CCA     | 24.14   | 19.71   | 21.93      | 33.16   | 31.66   | 32.41      |
| SM      | 22.64   | 21.84   | 22.24      | 36.85   | 38.67   | 37.76      |
| SCM     | 26.62   | 22.57   | 24.59      | 37.48   | 39.26   | 38.37      |
| GMMFA   | 23.09   | 20.34   | 21.72      | 28.41   | 24.87   | 26.64      |
| GMLDA   | 24.64   | 19.52   | 22.08      | 30.03   | 28.06   | 29.05      |
| JLSLR   | 23.60   | 21.22   | 22.41      | 39.42   | 36.91   | 38.17      |
| MDCR    | 26.19   | 21.03   | 23.61      | 41.07   | 37.75   | 39.41      |
| ACMR    | 33.22   | 24.50   | 28.86      | 50.61   | 42.82   | 46.72      |
| SGCR    | 28.42   | 22.71   | 25.57      | 43.65   | 40.60   | 42.10      |
| MRRDC   | 28.93   | 26.29   | 27.61      | 52.34   | 53.00   | 53.57      |

![Figure 3: Continued.](image-url)
Figure 3: Continued.
Figure 3: Precision-recall curves of the proposed MRRDC and compared methods. (a) I2T on wiki-R. (b) T2I on wiki-R. (c) I2T on wiki-W. (d) T2I on wiki-W. (e) I2T on Pascal. (f) T2I on Pascal. (g) I2T on INRIA-Websearch. (h) T2I on INRIA-Websearch.

Figure 4: Continued.
Figure 4: The mAP performance for each class on Wikipedia datasets and Pascal Sentence dataset. (a) I2I on wiki-R. (b) T2I on wiki-R. (c) Average on wiki-R. (d) I2I on wiki-W. (e) T2I on wiki-W. (f) Average on wiki-W. (g) I2I on Pascal. (h) T2I on Pascal. (i) Average on Pascal.
Figure 5: Class-level confusion matrices computed by the (a) image query and (b) text query on wiki-W.

Figure 6: Image queries and the corresponding top retrieved images, and text queries and the corresponding top retrieved images.
(1) and unlabelled data of all the different modalities are explored. The labelled information can be propagated to the unlabelled data during the training process.

(2) In most cases, GMLDA and GMMFA outperform CCA since GMLDA and GMMFA add category information to their formulation, which makes the common projection subspace more suitable for cross-modal retrieval.

(3) Compared with the shallow features, CNN features have great advantages for the I2T task, which is because CNN features can easily obtain the semantic information from original images directly.

To further verify the effectiveness of our proposed MRRDC, we also provide the confusion matrices on single-modal retrieval and the query examples for I2T and T2I in Figures 5 and 6 separately. Intuitively, from Figure 5, our method can achieve high precision in each category, which proves that the projection space is discriminative. We also observe from Figure 6 that, in many categories, our proposed method always successfully obtains the best retrieval results from query samples.

4.6. Convergence. Our objective formulation is solved by an iterative optimization algorithm. In a practical application, a fast retrieval speed is necessary. In Figure 7, we plot the convergence curves of our optimization algorithm as to the objective function value of equations (5) and (6) at each iteration on wiki-W and Pascal Sentence datasets separately. In this figure, the curve is monotonic at each iteration, and the algorithm generally converges within about 20 iterations for these datasets. The fast speed can ensure the high efficiency of our method.

5. Conclusion

In this paper, we propose an effective semi-supervised cross-modal retrieval approach based on discriminative mapping. Our approach uses different couples of discriminative projection matrices to map different modalities to the
common space where the correlation between different modalities can be maximum for different retrieval tasks. In particular, we use labelled samples to propagate the category information to unlabelled samples, and the original class information is preserved by using linear discriminant analysis. Therefore, the proposed method not only uses the relationship of different retrieval tasks but also keeps the structure information for different modalities. In the future, we will mine the correlation between different modalities and focus on the unsupervised cross-modal retrieval method for unlabelled data.

Data Availability

The data supporting this paper are from the reported studies and datasets in the cited references.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (no. 61702310), the Major Fundamental Research Project of Shandong, China (no. ZR2019ZD03), and the Taishan Scholar Project of Shandong, China (no. ts20190924).

References

[1] R. Bekkerman and J. Jeon, “Multi-modal clustering for multimedia collections,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, Minneapolis, MN, USA, July 2007.

[2] D. Eynard, A. Kovnatsky, M. M. Bronstein, K. Gashoff, and A. M. Bronstein, “Multimodal manifold analysis by simultaneous diagonalization of laplacians,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 12, pp. 2505–2517, 2015.

[3] S. Escalera, J. Gonzalez, X. Baro, and J. Shotton, “Guest editors’ introduction to the special issue on multimodal human pose recovery and behavior analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1489–1491, 2016.

[4] L. Liu, B. Zhang, H. Zhang, and N. Zhang, “Graph steered discriminative projections based on collaborative representation for image recognition,” Multimedia Tools and Applications, vol. 78, no. 17, pp. 24501–24518, 2019.

[5] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli, “MMALFM,” ACM Transactions on Information Systems, vol. 37, no. 2, pp. 1–28, 2019.

[6] 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.

[7] N. Rasiwasia, J. C. Pereira, E. Coviello et al., “A new approach to cross-modal multimedia retrieval,” in Proceedings of the International Conference on Multimedia-MM’10, pp. 251–260, Firenze, Italy, October 2010.

[8] L. Liu, S. Chen, X. Chen, T. Wang, and L. Zhang, “Fuzzy weighted sparse reconstruction error-steered semi-supervised learning for face recognition,” The Visual Computer, vol. 3, pp. 1–14, 2019.

[9] L. Zhu, Z. Huang, Z. Li, L. Xie, and H. T. Shen, “Exploring auxiliary context: discrete semantic transfer hashing for scalable image retrieval,” IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 11, pp. 5264–5276, 2018.

[10] L. Zhu, Z. Huang, X. Liu, X. He, J. Sun, and X. Zhou, “Discrete multimodal hashing with canonical views for robust mobile landmark search,” IEEE Transactions on Multimedia, vol. 19, no. 9, pp. 2066–2079, 2017.

[11] L. Zhu, X. Lu, Z. Cheng, J. Li, and H. Zhang, “Flexible multimodal hashing for scalable multimedia retrieval,” ACM Transactions on Intelligent Systems and Technology, vol. 11, no. 2, pp. 1–20, 2020.

[12] X. Lu, L. Zhu, Z. Cheng, X. Song, and H. Zhang, “Efficient discrete latent semantic hashing for scalable cross-modal retrieval,” Signal Processing, vol. 154, pp. 217–231, 2019.

[13] Y. Fang, H. Zhang, and Y. Ren, “Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing,” Knowledge-Based Systems, vol. 171, pp. 69–80, 2019.

[14] F. Shang, H. Zhang, L. J. Sun, and H. Zhang, “Adversarial cross-modal retrieval based on dictionary learning,” Neurocomputing, vol. 355, pp. 93–104, 2019.

[15] F. Shang, H. Zhang, J. Liu, and H. Zhang, “Semantic consistency cross-modal dictionary learning with rank constraint,” Journal of Visual Communication and Image Representation, vol. 62, pp. 259–266, 2019.

[16] M. Zhang, J. Li, H. Zhang, and L. Liu, “Deep semantic cross modal hashing with correlation alignment,” Neurocomputing, vol. 381, pp. 240–251, 2020.

[17] R. Rosipal and N. Kramer, “Overview and recent advances in partial least squares,” in Subspace, Latent Structure and Feature Selection, Statistical and Optimization, pp. 34–51, Springer, Berlin, Germany, 2005.

[18] A. Sharma and D. W. Jacobs, “Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600, Providence, RI, USA, June 2011.

[19] Y. Chen, L. Wang, W. Wang, and Z. Zhang, “Continuum regression for cross-modal multimedia retrieval,” in Proceedings of the IEEE International Conference on Image Processing, pp. 1949–1952, Orlando, FL, USA, September 2012.

[20] J. B. Tenenbaum and W. T. Freeman, “Separating style and content with bilinear models,” Neural Computation, vol. 12, no. 6, pp. 1247–1283, 2000.

[21] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, “Generalized multiview analysis: a discriminative latent space,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2160–2167, Providence, RI, USA, June 2012.

[22] V. Mahadevan, C. W. Wong, J. C. Pereira et al., “Maximum covariance unfolding: manifold learning for bimodal data,” in Advances in Neural Information Processing Systems, pp. 918–926, 2011.

[23] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, “Parallel field alignment for cross media retrieval,” in Proceedings of the ACM Multimedia Conference, pp. 897–906, Dallas, Texas, USA, April 2013.
(24) D. Lin and X. Tang, “Inter-modality face recognition,” in European Conference on Computer Vision, pp. 13–26, Springer, Berlin, Germany, 2006.

(25) 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.

(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, 2019.

(27) X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, “Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval,” IEEE Transactions on Cybernetics, vol. 50, no. 6, pp. 2400–2413, 2020.

(28) 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.

(29) Y. Peng and J. Qi, “CM-GANs: Cross-modal generative adversarial networks for common representation learning,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 15, no. 1, pp. 1–24, 2019.

(30) Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, “Supervised coupled dictionary learning with group structures for multi-modal retrieval,” in Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1070–1076, Bellevue, WA, USA, July 2013.

(31) X. Xu, A. Shimada, R. Taniguchi, and L. He, “Coupled dictionary learning and feature mapping for cross-modal retrieval,” in Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 1–6, Turin, Italy, June 2015.

(32) C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, “Discriminative dictionary learning with common label alignment for cross-modal retrieval,” IEEE Transactions on Multimedia, vol. 18, no. 2, pp. 208–218, 2016.

(33) Y. Wei, Y. Zhao, Z. Zhu et al., “Modality-dependent cross-media retrieval,” ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 4, pp. 1–13, 2016.

(34) 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, Tokyo, Japan, August 2017.

(35) B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in Proceedings of the 2017 ACM on Multimedia Conference-MM’17, pp. 154–162, Mountain View, CA, USA, October 2017.

(36) G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, “Generalized multi-view embedding for visual recognition and cross-modal retrieval,” IEEE Transactions on Cybernetics, vol. 48, no. 9, pp. 2542–2555, 2018.

(37) 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,” Journal of Visual Communication and Image Representation, vol. 58, pp. 1–11, 2019.

(38) Y. Ke and R. Sukthankar, “PCA-SIFT: a more distinctive representation for local image descriptors,” in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 506–513, Washington, DC, USA, July 2004.

(39) D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.