Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval

Kaiye Wang, Ran He, *Senior Member, IEEE*, Liang Wang, *Senior Member, IEEE*, Wei Wang, and Tieniu Tan, *Fellow, IEEE*

Abstract—Cross-modal retrieval has recently drawn much attention due to the widespread existence of multimodal data. It takes one type of data as the query to retrieve relevant data objects of another type, and generally involves two basic problems: the measure of relevance and coupled feature selection. Most previous methods just focus on solving the first problem. In this paper, we aim to deal with both problems in a novel joint learning framework. To address the first problem, we learn projection matrices to map multimodal data into a common subspace, in which the similarity between different modalities of data can be measured. In the learning procedure, the ℓ_{21}-norm penalties are imposed on the projection matrices separately to solve the second problem, which selects relevant and discriminative features from different feature spaces simultaneously. A multimodal graph regularization term is further imposed on the projected data,which preserves the inter-modality and intra-modality similarity relationships.An iterative algorithm is presented to solve the proposed joint learning problem, along with its convergence analysis. Experimental results on cross-modal retrieval tasks demonstrate that the proposed method outperforms the state-of-the-art subspace approaches.

Index Terms—Subspace learning, coupled feature selection, half-quadratic minimization, cross-modal retrieval

1 Introduction

In some real applications, data are often represented in different ways or collected from diverse domains. As a result, the data related to the same underlying content or object may exist in different modalities and exhibit heterogeneous properties. For example, when visiting the Great Wall, we may record it by taking pictures, posting a piece of microblog, or recording a video clip. These data present the same content, but they take different forms. With the rapid growth of such multimedia data, there is an immediate need for efficiently and effectively analyzing the data across different modalities. Although much attention has been paid to multimodal data analysis [1], [2], [3], [4], the common strategy is to integrate multiple modalities to improve the learning performance. In this paper, we focus on cross-modal retrieval, which aims to take one type of data as query to retrieve relevant data objects of another type. For example, a user can use a text to retrieve relevant pictures and videos, or search relevant textual descriptions or videos by submitting an interesting image as a query. Cross-modal retrieval enables users to take any modality of content at hand as a query. The search results of the cross-modal retrieval are rich in multiple modalities, and thus are more comprehensive than the results of traditional single-modality retrieval methods.

The cross-modal retrieval generally involves two basic problems: the measure of relevance and coupled feature selection. As shown in the next section, although some methods have been proposed to solve the cross-modal problem, most of them learn a common latent subspace to make all modalities of data comparable. They mainly solve the first basic problem, but the second one is not well addressed, that is, how to simultaneously select relevant and discriminative features from multimodal feature spaces. Here we call the second problem "coupled feature selection". Although various feature selection methods [5], [6] have been developed for the single modality data analysis, few previous work attempts to perform common subspace learning and feature selection simultaneously, which have been shown to be coupled problems for the cross-modal problem [7].

To solve both problems mentioned above, this paper proposes a novel joint learning framework (as shown in Fig. 1) for the cross-modal retrieval problem by combining common subspace learning and coupled feature selection. Firstly, inspired by the potential relationship between Canonical Correlation Analysis (CCA) and linear least squares [8], coupled linear regression is used to learn projection matrices to map data from different modalities into a common subspace. At the same time, ℓ_{21}-norm is used to select the relevant and discriminative features from different feature spaces simultaneously, and a multimodal graph regularization is used to preserve the inter-modality and intra-modality similarity relationships when mapping. Secondly, based on the half-quadratic analysis for ℓ_{21}-norm [9], we develop an iterative algorithm to solve the proposed joint learning problem, and prove its convergence. Finally, the proposed method is applied to several cross-modal retrieval tasks. Experimental results on three publicly available datasets show that the proposed method outperforms previous subspace approaches.

• *The authors are with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China. E-mail: {kaiye.wang, rhe, wangliang, wangwei, tnt}@nlpr.ia.ac.cn.*

Manuscript received 23 July 2014; revised 30 Apr. 2015; accepted 17 Nov. 2015. Date of publication 2 Dec. 2015; date of current version 12 Sept. 2016. Recommended for acceptance by G. Mori.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2505311

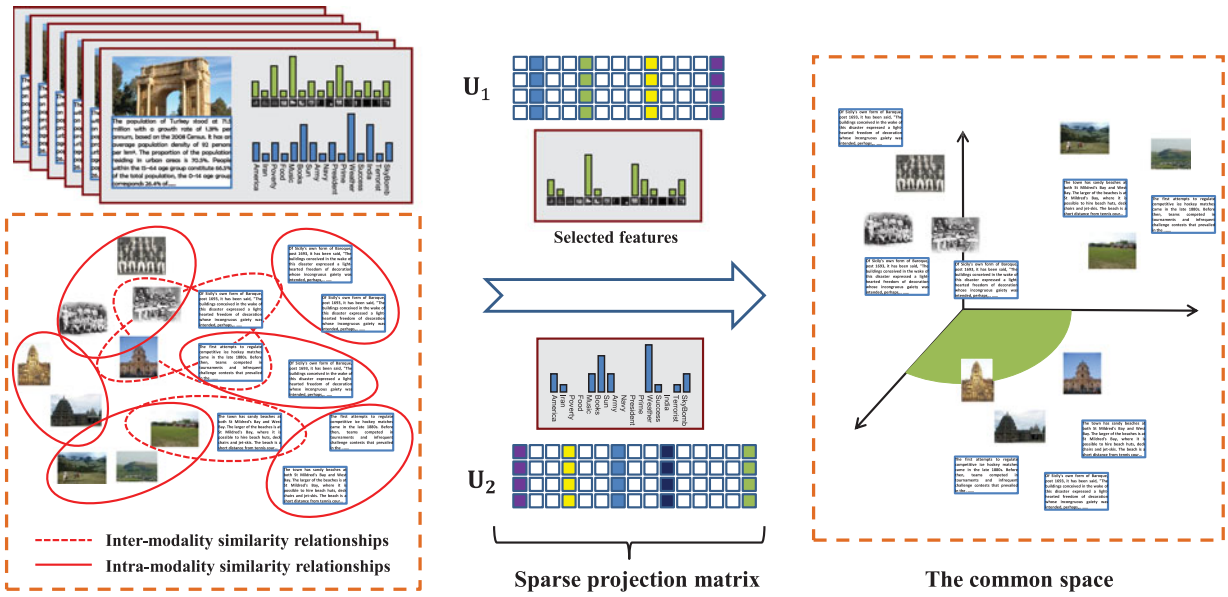


Fig. 1. The overview of the proposed method. U_1 and U_2 are projection matrices learned by our method for image and text spaces, respectively. They project different modalities of data into a common space, while performing feature selection on different feature spaces. The proposed method also preserves the inter-modality and intra-modality similarity relationships when mapping.

The main contributions of our work include the following:

- The proposed joint learning method elegantly combines common subspace learning and coupled feature selection into a single framework. Experimental results on cross-modal retrieval tasks show its superiority.
- When learning projection matrices, a multimodal graph regularization term is proposed to explicitly preserve the inter-modality and intra-modality similarity among multimodal data objects, which further improves the performance.
- An iterative algorithm based on half-quadratic minimization is proposed to efficiently solve the joint learning objective function. Experimental results demonstrate that it obtains promising results and performs better than the state-of-the-art subspace methods.

The remainder of this paper is organized as follows. In Section 2, we briefly overview related work on the cross-modal retrieval problem. Section 3 describes our proposed joint learning framework for cross-modal retrieval, along with an iterative algorithm to solve this problem. In Section 4, we report experimental results on several cross-modal datasets. Finally, we conclude the paper in Section 5.

2 Related Work

Since the cross-modal retrieval is considered as an important problem in some real applications, various approaches have been proposed to deal with this problem, such as probabilistic models, metric learning approaches and subspace learning methods, which will be introduced as follows, respectively.

2.1 Probabilistic Models

Probabilistic models have been widely applied to a specific cross-modal problem, i.e., image annotation [10], [11]. To capture the correlation between images and annotations, Latent Dirichlet Allocation (LDA) [12] has been extended to learn the joint distribution of multi-modal data such as Correspondence LDA (Corr-LDA) [10] and Topic-regression Multi-modal LDA (Tr-mm LDA) [11]. Corr-LDA uses topics as the shared latent variables, which represent the underlying causes of cross-correlations in the multi-modal data. Tr-mm LDA learns two separate sets of hidden topics and a regression module which captures more general forms of association and allows one set of topics to be linearly predicted from the other. Jia et al. [13] propose a new probabilistic model to learn a set of shared topics across the modalities. The model defines a Markov random field on the document level which allows modeling more flexible document similarities. Srivastava and Salakhutdinov propose a deep Boltzmann machine [14] to learn a multimodal representation for multimodal data. However, these aforementioned approaches just tend to model the one-to-one correlation, which may not be able to capture the complex structure of the multimodal data.

2.2 Metric Learning Approaches

Another perspective for cross-modal retrieval is to learn a metric between different modalities of data. Li et al. [15] introduce a cross-modal factor analysis (CFA) approach to evaluate the association between two modalities. The CFA method adopts a criterion of minimizing the Frobenius norm between pairwise data in the transformed domain. Wu et al. [16] study the metric learning problem to find the similarity function over two different spaces. Mignon and Jurie [17] propose a metric learning approach for cross-modal matching, which considers both positive and negative constraints. Quadrianto and Lampert [18] propose a new metric learning scheme to learn projections from the data in different modalities into a shared feature space, in which the Euclidean distance provides a meaningful intra-modality and inter-modality similarity. Zhai et al. [19]

propose a regularized metric learning algorithm to learn a heterogeneous metric for cross media retrieval. Lu et al. [20] and Wu et al. [21] study the cross-modal retrieval as a problem of learning to rank. They utilize the structural SVM to learn a metric such that the ranking of the data induced by the distance from a query can be optimized against various ranking measures. In [22], [23], [24], the learnt Hamming metric is used to speed up the cross-modal search, but the Hamming metric is discrete-valued so that its retrieval accuracy is lower. These methods mentioned above generally treat similar pairs and dissimilar pairs or rank lists equally when modeling the structure of the multimodal data. However, some less informative pairs and rank lists may potentially lead the model to depart from the correct structure, which degrades the performance.

2.3 Subspace Learning Methods

Recently, several approaches for establishing inter-modal relationships between data from different modalities generally rely on subspace learning, such as Canonical Correlation Analysis [25], [26], Partial Least Squares (PLS) [27] and Bilinear Model (BLM) [28], [29]. Specifically, CCA is probably the most popular one due to its wide-spread use in cross-media retrieval [30], [31], [32], cross-lingual retrieval [33] and some vision problems [34].

Rasiwasia et al. [30] address the cross-modal retrieval problem by investigating the correlations between two modalities, where CCA is proved to be effective. Li et al. [34] apply CCA to face recognition based on non-corresponding region matching. They use CCA to learn a common space in which the possibility of whether two non-corresponding face regions belong to the same face can be measured. Recently, Partial Least Squares [27] is also used for the cross-modal matching problem. To perform multimodal face recognition, such as front versus profile, photos vs. sketches, and high-resolution photos versus low-resolution photos, Sharma and Jacobs [35] use PLS to linearly map images in different modalities to a common linear subspace in which they are highly correlated. Chen et al. [36] apply PLS to the cross-modal document retrieval. They use PLS to switch the image features into the text space, then learn a semantic space for the measure of similarity between two different modalities. In [29], Tenenbaum and Freeman propose a bilinear model to derive a common space for cross-modal face recognition, and BLM is also used for text-image retrieval in [28].

Besides CCA, PLS and BLM, there are some other methods for the cross-modal problem. Mahadevan et al. [37] propose maximum covariance unfolding, a manifold learning algorithm for simultaneous dimensionality reduction of data from different input modalities. Mao et al. [38] introduce a method for cross media retrieval, named parallel field alignment retrieval, which integrates a manifold alignment framework form the perspective of vector fields. Lin and Tang [39] propose a common discriminant feature extraction (CDFE) method to learn a common feature subspace where the difference of within scatter matrix and between scatter matrix is maximized. Recently, Sharma et al. [28] extend Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA) to their multiview counterparts, i.e., Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA), and apply them to deal with the cross-media retrieval problem. GMLDA and GMMFA take the semantic category into account, and obtain promising results.

Besides these three kinds of methods, there are some other methods [40], [41], [42], [43], [44], [45] proposed for the cross-modal problem, such as the dictionary learning method [40], the graph-based learning method [42], the constraint propagation method [43], [44], and so on. There is also some related work on multi-view embedding methods [46], [47], [48], [49], [50], which have been applied to document categorization, image annotation, multi-class classification, etc. Weston et al. [48] propose to learn a joint space for images and annotations and optimize a ranking objective function to rank the annotations for images. However, the ranking error functions are explicitly designed for annotations. It cannot be directly used in general multimodal case. The algorithm in [49] maps classes into a semantic space, and then maps the document into the same space for document categorization. To deal with large multi-class tasks, Bengio et al. [46] propose to embed labels into a low dimensional space. Since the classes/labels are supervised information, these multi-view methods are actually developed for dealing with one modality of data.

Thanks to the continuous effort made by researchers, we have witnessed great advances in the cross-modal retrieval field. However, most of them mainly focus on the measurement of relevance, and coupled feature selection has not been well addressed. Since the dimensionality of real world data is often high, there are naturally redundant and irrelevant features. Hence, how to simultaneously select the relevant and discriminative features for different modalities of data is very important. Accordingly, we aim to jointly perform common subspace learning and coupled feature selection in this work. To achieve this goal, we propose a generic minimization formulation by combining linear regressions, ℓ_{21}-norms and a multimodal graph regularization term, which will be detailed in the next section.

This paper is built upon our preliminary conference version [7]. The main extensions are summarized as follows.

(1) While the previous method in [7] applied the low-rank constraint to enhance the relevance of similar objects, we now propose a multimodal graph to better model the similarity relationships among different modalities of data, which is demonstrated to outperform the low-rank constraint in terms of both computational cost and retrieval performance.
(2) Accordingly, a new iterative algorithm is proposed to solve the modified objective function and the proof of its convergence is given. Furthermore, we validate the convergence of this iterative algorithm in our experiments.
(3) Experimentally, we add new experiments on the NUS-WIDE dataset, which further validate the effectiveness of our method. In addition, we provide additional discussion of the experimental results, and analyze the parameters sensitivity of the proposed method.

3 Joint Feature Selection and Subspace Learning

3.1 Problem Formulation

We begin with a brief introduction to some notations used here. For matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its ith row, jth column are denoted by $\mathbf{m}^i$, $\mathbf{m}_j$ respectively. The Frobenius norm of the matrix $\mathbf{M}$ is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{n} \|\mathbf{m}^i\|_2^2}$. And $\|\mathbf{M}\|_{2,1}$ is the sum of the ℓ_2-norm of the rows of $\mathbf{M}$: $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \|\mathbf{m}^i\|_2$.

3.1.1 Problem Statement

Suppose that we have a collection of data from M different modalities, with $\{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^M\}$ representing the same underlying content or objects. For example, the text, image, audio and video are often used to describe the same topic. Given a query from one modality, the goal of the cross-modal retrieval is to return the top k closest matches in another modality.

As mentioned above, the cross-modal retrieval generally involves two problems. Previous methods mainly focus on the first problem. They project data from different modalities into a latent space, in which the possibility of whether two different modal data represent the same semantic concept can be measured. However, the second problem, i.e., coupled feature selection, is usually ignored. Based on this consideration, we propose that the feature selection procedure should be performed on different feature spaces simultaneously for better retrieval.

3.1.2 The Objective Function

Let $\mathcal{L} = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^M\}_{i=1}^N$ denote N labeled multimodal documents, and each document contains data from M different modalities representing the same semantic. Since the data from different modalities $\mathcal{L}_i = \{\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^M\}$ represent the same semantic, they should share the same representation in a common space, denoted by $\mathbf{y}_i$. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]^T$ denote the representation matrix. $\mathbf{X}_p = [\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_N^p] \in \mathbb{R}^{d_p\times N}, p = 1, 2, \ldots, M$ represent the labeled data matrices from M modalities, respectively. $\mathbf{X}_p^b = [\mathbf{x}_1^p, \mathbf{x}_2^p, \ldots, \mathbf{x}_{N+E_p}^p] \in \mathbb{R}^{d_p\times(N+E_p)}, p = 1, 2, \ldots, M,$ represent the matrices of both labeled and unlabeled data. The pth modality has N labeled samples and E_p unlabeled samples embedded in the d_p dimensional space. The goal of our method is to learn a projection matrix for each modality of data, which can be used to project data from different modalities into a common space. Then, we can perform cross-modal retrieval in the common space.

Firstly, we learn the low-dimensional representations $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]^T$ for the multimodal documents $\{\mathcal{L}_i\}_{i=1}^N$. In the general graph embedding framework, the optimal embedding $\mathbf{Y}$ can be obtained by

$$\min_{\mathbf{Y}} \sum_{i,j} w_{ij} ||\mathbf{y}_i - \mathbf{y}_j||^2 \tag{1}$$

under the constraint $\mathbf{Y}^T\mathbf{B}\mathbf{Y} = I$, and $\mathbf{B}$ is a diagonal matrix with $\mathbf{B}_{ii} = \sum_j w_{ij}$. w_{ij} indicates whether the multimodal documents $\mathcal{L}_i$ and $\mathcal{L}_j$ represent similar semantic, which is defined as follows:

$$w_{ij} = \begin{cases} 1/N_t, & \text{if } \mathcal{L}_i \text{ and } \mathcal{L}_j \text{ belong to the } t\text{th class} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where N_t is the number of documents in the tth class. The solution of Eq. (1) is given as below [51]:

$$\mathbf{v}_t = (\underbrace{0, \ldots 0}_{\sum_{i=1}^{t-1} N_i}, \underbrace{1, \ldots 1}_{N_t}, \underbrace{0, \ldots, 0}_{\sum_{i=t+1}^{c} N_i}), t = 1, \ldots, c, \tag{3}$$

where c is the number of classes. So the low-dimensional representations are obtained as $\mathbf{Y} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c) \in \mathbb{R}^{N\times c}$.

Next, we learn a projection matrix for each modality to map different types of data into the common space, in which the similarity between different modalities of data can be measured. At the same time, we perform ℓ_{21}-norm on the projection matrices for coupled feature selection, and preserve the inter-modality and intra-modality similarity relationships among M different modalities of data objects. That is, we have a generic minimization problem in the following form,

$$\begin{aligned} \min_{\mathbf{U}_1,\ldots,\mathbf{U}_M} & \sum_{p=1}^{M} ||\mathbf{X}_p^T\mathbf{U}_p - \mathbf{Y}||_F^2 + \lambda_1 \sum_{p=1}^{M} ||\mathbf{U}_p||_{21} \\ & + \lambda_2 \Omega(\mathbf{U}_1, \ldots, \mathbf{U}_M), \end{aligned} \tag{4}$$

where $\mathbf{U}_p, p = 1, \ldots, M$ are the projection matrices for the M modalities of data respectively. The first term is a coupled linear regression term, which is used to learn projection matrices for mapping different modal data into the common space. The second term contains M ℓ_{21}-norms that play a role of feature selection on different feature spaces simultaneously. The third term is a multimodal graph regularization, which can preserve the inter-modality and intra-modality similarity relationships effectively. Note that the third term is different from that in the previous version [7]. The multimodal graph regularization is more helpful for modeling the structure of the multimodal data.

3.1.3 The Multimodal Graph Regularization

We use both labeled and unlabeled data to construct the multimodal graph according to two kind of relationships as shown in Fig. 2. The two kinds of relationships among the multimodal data are defined as follows:

1) *Inter-modality similarity relationship*: it is well known that although different modalities of data have different representations and are in different feature spaces, they share similar semantics if they are related to the same content or topic, which can be understood as the inter-modality similarity relationship. For example, if they belong to the same class, they are similar on the topic. We hope to preserve the inter-modality similarity relationship when learning the common space. According to the inter-modality similarity relationship, we define the similarity matrix W^{pq} between the pth modality and the qth modality as follows:

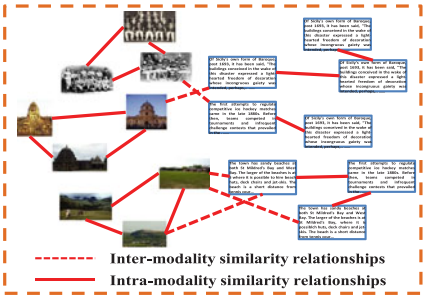


Fig. 2. The multimodal graph is constructed according to inter-modality and intra-modality similarity relationships.

$$W_{ij}^{pq} = \begin{cases} 1, & \text{if } \mathbf{x}_i^p \text{ has similar semantics to } \mathbf{x}_j^q \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

2) *Intra-modality similarity relationship*: we also hope to preserve the intra-modality similarity relationship among data objects within each single modality, i.e., the data objects with the neighborhood relationship should be close to each other in the common space. To preserve the local structural information within each single modality, a kNN similarity graph is constructed here. The similarity matrix W^p within the pth modality is defined as below:

$$W_{ij}^p = \begin{cases} \exp(-z_{ij}^p/2\sigma^2), & \text{if } \mathbf{x}_i^p \in N_k(\mathbf{x}_j^p) \text{ or } \mathbf{x}_j^p \in N_k(\mathbf{x}_i^p) \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where z_{ij}^p is the Euclidean distance between $\mathbf{x}_i^p$ and $\mathbf{x}_j^p$, i.e., $z_{ij}^p = \left\|\mathbf{x}_i^p - \mathbf{x}_j^p\right\|^2$, and $N_k(\mathbf{x}_i^p)$ denotes the set of k nearest neighbors of $\mathbf{x}_i^p$.

According to the two kinds of similarity relationships, we feed all different modalities of data into a joint multimodal graph. The overall similarity matrix W is defined as follows:

$$\mathbf{W} = \begin{bmatrix} \beta\mathbf{W}^1 & \mathbf{W}^{12} & \dots & \mathbf{W}^{1M} \\ \mathbf{W}^{21} & \beta\mathbf{W}^2 & \dots & \mathbf{W}^{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}^{M1} & \mathbf{W}^{M2} & \dots & \beta\mathbf{W}^M \end{bmatrix}, \tag{7}$$

where β is a parameter which balances the effect of the inter-modality similarity and the intra-modality similarity. $\mathbf{W}^{ij}, i,j = 1,\dots,M$ indicates the inter-modality similarity defined by Eq. (5), and $\mathbf{W}^i, i = 1,\dots,M$ indicates the intra-modality similarity defined by Eq. (6).

Based on the multimodal graph, the third term of Eq. (4) is defined as follows:

$$\begin{aligned} \Omega(\mathbf{U}_1,...,\mathbf{U}_M) &= \frac{1}{2}\sum_{i=1}^{\hat{N}}\sum_{j=1}^{\hat{N}} W_{ij}||f_i - f_j||^2 \\ &= Tr(\mathbf{FLF}^T), \end{aligned} \tag{8}$$

where $\hat{N}$ is the number of the total samples from all modalities, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix. $\mathbf{F} = (\mathbf{F}_1^T,\dots,\mathbf{F}_M^T) = (\mathbf{U}_1^T\mathbf{X}_1^b,\dots,\mathbf{U}_M^T\mathbf{X}_M^b)$ denotes all modalities of projected data in the common space. Eq. (8) can be rewritten as

$$\Omega(\mathbf{U}_1,...,\mathbf{U}_M) = \sum_{p=1}^{M}\sum_{q=1}^{M} Tr(\mathbf{U}_p^T\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T\mathbf{U}_q). \tag{9}$$

The multimodal graph regularization term encourages a mapping which preserves the inter-modality and intra-modality similarity. It is a generalized graph which takes different kinds of data into consideration.

3.2 Optimization Algorithm

3.2.1 An Iterative Algorithm

In this subsection, an iterative algorithm based on the half-quadratic minimization [9], [52] is proposed to solve the objective function in (4).

If we define $\phi(x) = \sqrt{x^2+\varepsilon}$, we can replace $\left\|\mathbf{U}_p\right\|_{21}$ with $\sum_{i=1}^{d_p}\phi(\left\|\mathbf{u}_p^i\right\|_2)$, according to the analysis for the ℓ_{21}-norm in [9]. And ε is a smoothing term, which is usually set to be a small value. It can be proved that $\phi(x) = \sqrt{x^2+\varepsilon}$ satisfies all conditions as follows:

$$\begin{aligned} x \to &\phi(x) \text{ is convex on } R, \\ x \to &\phi(\sqrt{x}) \text{ is concave on } R_+, \\ \phi(x) = &\phi(-x), \forall x \in R, \\ &\phi(x) \text{ is } C^1 \text{on} R, \\ \phi'(0^+) > &0, \lim_{x\to\infty}\phi(x)/x^2 = 0. \end{aligned} \tag{10}$$

Then, we can optimize $\phi(\cdot)$ in a half-quadratic way [53] according to the following Lemma 1 [9].

Lemma 1. Let $\phi(\cdot)$ be a function satisfying all conditions in (10), for a fixed $\|\mathbf{u}^i\|_2$, there exists a dual potential function $\varphi(\cdot)$, such that

$$\phi(\left\|\mathbf{u}^i\right\|_2) = \inf_{s\in R}\left\{s\left\|\mathbf{u}^i\right\|_2^2 + \varphi(s)\right\}, \tag{11}$$

where s is determined by the minimizer function $\varphi(\cdot)$ with respect to $\phi(\cdot)$.

According to Lemma 1, the objective function in (4) can be reformulated as follows:

$$\begin{aligned} \min_{\mathbf{U}_1,\dots,\mathbf{U}_M} &\sum_{p=1}^{M}||\mathbf{X}_p^T\mathbf{U}_p - \mathbf{Y}||_F^2 + \lambda_1\sum_{p=1}^{M}Tr(\mathbf{U}_p^T\mathbf{R}_p\mathbf{U}_p) \\ &+ \lambda_2\sum_{p=1}^{M}\sum_{q=1}^{M}Tr(\mathbf{U}_p^T\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T\mathbf{U}_q), \end{aligned} \tag{12}$$

where $\mathbf{R}_p = Diag(\mathbf{r}_p)$. And $\mathbf{r}_p$ is an auxiliary vector of the ℓ_{21}-norm, where the ith element $r_p^i = 12||\mathbf{u}_p^i||_2$. The elements of $\mathbf{r}_p$ are regularized respectively as follows[1]:

$$r_p^i = \frac{1}{2\sqrt{||\mathbf{u}_p^i||_2^2 + \varepsilon}}, \tag{13}$$

1. Note that $||\mathbf{u}_p^i||_2$ can be zero theoretically. However, we cannot set r_p^i to zero, otherwise the iterative algorithm cannot be guaranteed to converge. To solve this problem, we regularize r_p^i in Eq. (13).

where ε is a smoothing term, which is usually set to be a small constant value.

Differentiating the objective function in (12) with respect to $\mathbf{U}_p$ and setting it to zero, we have the following equation:

$$\begin{aligned}\mathbf{X}_p(\mathbf{X}_p^T\mathbf{U}_p-\mathbf{Y})+\lambda_1\mathbf{R}_p\mathbf{U}_p+\lambda_2\mathbf{X}_p^b\mathbf{L}_{pp}(\mathbf{X}_p^b)^T\mathbf{U}_p\\+\lambda_2\sum_{q\neq p}\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T\mathbf{U}_q=0,\end{aligned}\tag{14}$$

which can be rewritten as

$$\begin{aligned}&(\mathbf{X}_p\mathbf{X}_p^T+\lambda_1\mathbf{R}_p+\lambda_2\mathbf{X}_p^b\mathbf{L}_{pp}(\mathbf{X}_p^b)^T)\mathbf{U}_p\\&=\mathbf{X}_p\mathbf{Y}-\lambda_2\sum_{q\neq p}\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T\mathbf{U}_q.\end{aligned}\tag{15}$$

Then, the optimal solution of (14) can be computed via solving the above linear system problem.

Algorithm 1: Joint Feature Selection and Subspace Learning (JFSSL)

Input:

The matrix of both labeled and unlabeled data $\mathbf{X}_p^b \in \mathbb{R}^{d_p\times(N+E_p)}$;

The matrix of labeled data $\mathbf{X}_p\in\mathbb{R}^{d_p\times N}$;

The low-dim representation $\mathbf{Y}\in\mathbb{R}^{N\times c}$;

Output:

The projection matrices $\mathbf{U}_p\in\mathbb{R}^{d_p\times c}, p=1,\ldots,M$;

(a) Compute the Laplacian matrix of the multimodal graph $\mathbf{L}$;

(b) Set $t=0$. Initialize $\mathbf{U}_p, p=1,\ldots,M$ as identity matrix.

repeat

1. Compute $\mathbf{r}_p^t$ according to Eq. (13).
2. By solving the linear system problem in Eq. (15), $\mathbf{U}_p^t$ is updated as follows:

$$\begin{aligned}\mathbf{U}_p^{t+1}=(\mathbf{X}_p\mathbf{X}_p^T+\lambda_1\mathbf{R}_p+\lambda_2\mathbf{X}_p^b\mathbf{L}_{pp}(\mathbf{X}_p^b)^T)^{-1}(\mathbf{X}_p\mathbf{Y}\\-\lambda_2\sum_{q\neq p}\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T\mathbf{U}_q^t)\end{aligned}\tag{16}$$

3. $t=t+1$

until Converges

Algorithm 1 summarizes the alternate minimization procedure to optimize (4). Firstly, we construct the multimodal graph and compute the graph Laplacian matrix $\mathbf{L}$ in Step (a); In Step 1 of the loop, we compute the auxiliary vectors $\mathbf{r}_p, p=1,\ldots,M$ that correspond to the ℓ_{21}-norms and play an important role in feature selection on different feature spaces. In Step 2, we find the solution $\mathbf{U}_p, p=1,\ldots,M$. Here the iteration continues until convergence. In our experiments, it takes about five iterations before convergence.

3.2.2 Convergence

We show that the proposed iterative algorithm in Algorithm 1 converges by the following theorem.

Theorem 1. The iterative algorithm in Algorithm 1 will monotonically decrease the objective function in Eq. (12) in each iteration until convergence.

Proof. The problem in Eq. (12) is equivalent to

$$\begin{aligned}\min_{\mathbf{U}}\ &Tr(\mathbf{U}^T\mathbf{R}\mathbf{U})+\gamma Tr(\mathbf{U}^T\mathbf{G}\mathbf{U})\\&s.t.\ \mathbf{X}_p^T\mathbf{U}_p=\mathbf{Y}, p=1,\ldots,M,\end{aligned}\tag{17}$$

where $\gamma=\lambda_2/\lambda_1$, $\mathbf{U}$, $\mathbf{R}$ and $\mathbf{G}$ are defined as follows:

$$\mathbf{U}=\begin{bmatrix}\mathbf{U}_1\\\mathbf{U}_2\\\vdots\\\mathbf{U}_M\end{bmatrix},\mathbf{R}=\begin{bmatrix}\mathbf{R}_1&&&\\&\mathbf{R}_2&&\\&&\ddots&\\&&&\mathbf{R}_M\end{bmatrix},\tag{18}$$

$$\mathbf{G}=\begin{bmatrix}\mathbf{G}_{11}&\mathbf{G}_{12}\ldots\mathbf{G}_{1M}\\\mathbf{G}_{21}&\mathbf{G}_{22}\ldots\mathbf{G}_{2M}\\\vdots&\vdots\ \ddots\vdots\\\mathbf{G}_{M1}&\mathbf{G}_{M2}\ldots\mathbf{G}_{MM}\end{bmatrix},\tag{19}$$

where we set $\mathbf{G}_{pq}=\mathbf{X}_p^b\mathbf{L}_{pq}(\mathbf{X}_q^b)^T,\ p,q=1,\ldots,M$.

According to Algorithm 1, it can be inferred from (17) that

$$\begin{aligned}\mathbf{U}_{t+1}=\min_{\mathbf{U}}&Tr(\mathbf{U}^T(\mathbf{R}_t+\gamma\mathbf{G})\mathbf{U})\\&s.t.\ \mathbf{X}_p^T\mathbf{U}_p=\mathbf{Y}, p=1,\ldots,M.\end{aligned}\tag{20}$$

Therefore, we have

$$\begin{aligned}&Tr(\mathbf{U}_{t+1}^T(\mathbf{R}_t+\gamma\mathbf{G})\mathbf{U}_{t+1})\leqslant Tr(\mathbf{U}_t^T(\mathbf{R}_t+\gamma\mathbf{G})\mathbf{U}_t)\\&\Rightarrow\sum_{i=1}^d\frac{\|\mathbf{u}_{t+1}^i\|_2^2}{2\|\mathbf{u}_t^i\|_2}+\gamma Tr(\mathbf{U}_{t+1}^T\mathbf{G}\mathbf{U}_{t+1})\\&\leqslant\sum_{i=1}^d\frac{\|\mathbf{u}_t^i\|_2^2}{2\|\mathbf{u}_t^i\|_2}+\gamma Tr(\mathbf{U}_t^T\mathbf{G}\mathbf{U}_t).\end{aligned}\tag{21}$$

Then we have the following inequality

$$\begin{aligned}&\sum_{i=1}^d\|\mathbf{u}_{t+1}^i\|_2+\gamma Tr(\mathbf{U}_{t+1}^T\mathbf{G}\mathbf{U}_{t+1})\\&-\left(\sum_{i=1}^d\|\mathbf{u}_{t+1}^i\|_2-\sum_{i=1}^d\frac{\|\mathbf{u}_{t+1}^i\|_2^2}{2\|\mathbf{u}_t^i\|_2}\right)\\&\leqslant\sum_{i=1}^d\|\mathbf{u}_t^i\|_2+\gamma Tr(\mathbf{U}_t^T\mathbf{G}\mathbf{U}_t)\\&-\left(\sum_{i=1}^d\|\mathbf{u}_t^i\|_2-\sum_{i=1}^d\frac{\|\mathbf{u}_t^i\|_2^2}{2\|\mathbf{u}_t^i\|_2}\right).\end{aligned}\tag{22}$$

It has been shown in [54] that for any nonzero vectors $\mathbf{u}_t^i|_{t=1}^h$, the following inequality holds:

$$\begin{aligned}&\sum_{i=1}^d\|\mathbf{u}_{t+1}^i\|_2-\sum_{i=1}^d\frac{\|\mathbf{u}_{t+1}^i\|_2^2}{2\|\mathbf{u}_t^i\|_2}\\&\leqslant\sum_{i=1}^d\|\mathbf{u}_t^i\|_2-\sum_{i=1}^d\frac{\|\mathbf{u}_t^i\|_2^2}{2\|\mathbf{u}_t^i\|_2},\end{aligned}\tag{23}$$

where h is an arbitrary number. Thus, we can get the following inequality:

$$\sum_{i=1}^{d} \left\|\mathbf{u}_{t+1}^{i}\right\|_2 + \gamma Tr(\mathbf{U}_{t+1}^{T}\mathbf{G}\mathbf{U}_{t+1}) \leqslant \sum_{i=1}^{d} \left\|\mathbf{u}_{t}^{i}\right\|_2 + \gamma Tr(\mathbf{U}_{t}^{T}\mathbf{G}\mathbf{U}_{t}), \tag{24}$$

which indicates that the objective function value of (12) monotonically decreases until convergence using the proposed iterative approach in Algorithm 1. □

3.2.3 Complexity

Finally, we briefly discuss the computational complexity. Let $N_p, p = 1, 2, \ldots, M$ denote the number of the samples belonging to the p-modality. For the multimodal graph, constructing the inter-modality similarity matrix takes $O(\frac{1}{2}\sum_{p=1}^{M}\sum_{q=1}^{M} N_p N_q)$ time, and constructing the intra-modality similarity matrix takes $O(\sum_{p=1}^{M}(d_p + k)N_p^2)$ time. The values of M and k are usually small constants. Hence, the complexity of computing the Laplacian matrix is approximately $O(d_m N_m^2)$, where $d_m N_m^2$ for the mth modality is the largest one among all modalities. In Step 2, instead of calculating the inverse of a few matrices, we update the projection matrices by solving a system of linear equations, among which the most complex part takes $O(\hat{d}^2)$ ($\hat{d} = \max(d_1, \ldots, d_p)$).

4 Experimental Results

4.1 Data Sets

The *Pascal VOC* dataset [55] consists of 5,011/4,952 (training/testing) image-tag pairs, which can be categorized into 20 different classes. Since some images are multi-labeled, we select images with only one object as the way in [28], resulting in 2,808 training and 2,841 testing data. The image features are 512-dimensional Gist features [55], and the text features are 399-dimensional word frequency features.

The dataset used here is the *NUS-WIDE* dataset [56]: each image is associated with user tags, which can be taken as an image-text pair. To guarantee that each class has abundant training samples like [57], we select those pairs that belong to one of the 21 largest classes with each pair exclusively belonging to one of the 21 classes, which results in 72,219 image-text pairs. The images are represented with a 500-dimensional SIFT feature vectors [58], and the textual tags are represented with 1,000-dimensional tag occurrence feature vectors. We take 50 percent of the data as the training set and the remaining as the testing set.

The *Wiki image-text* dataset [30], generated from Wikipedia's "featured article", consists of 2,866 image-text pairs. In each pair, the text is an article describing people, places or some events and the image is closely related to the content of the article. Each pair is labeled with one of 10 semantic classes. We split it into a training set of 1,300 pairs (130 pairs per class) and a testing set of 1,566 pairs. The representation of the text with 10 dimensions is derived from a latent Dirichlet allocation model [12]. The images are represented by the 128 dimensional SIFT descriptor histograms [58].

4.2 Evaluation Metrics

To evaluate the performance of the proposed JFSSL method, two cross-modal retrieval tasks are conducted: (1) Image query versus Text database, (2) Text query versus Image database. In testing phase, we map the multimodal data into the common subspace using the learned projection matrices. We take one modality of data of the testing set as the query set to retrieve another modality of data. The cosine distance is adopted to measure the similarity of features. Given an image (or text) query, the goal of each cross-modal task is to find the nearest neighbors from the text (or image) database.

The *mean average precision (MAP)* [30] is used to evaluate the overall performance of the tested algorithms. To compute MAP, we first evaluate the average precision (AP) of a set of R retrieved documents by $\mathrm{AP} = \frac{1}{T}\sum_{r=1}^{R} P(r)\delta(r)$ where T is the number of relevant documents in the retrieved set, $P(r)$ denotes the precision of the top r retrieved documents, and $\delta(r) = 1$ if the rth retrieved document is relevant (where relevant means belonging to the class of the query) and $\delta(r) = 0$ otherwise. The MAP is then computed by averaging the AP values over all queries in the query set. The larger the MAP, the better the performance.

Besides the MAP, we also use *precision-scope curve* [59] and *precision-recall curve* [30] to evaluate the effectiveness of different methods. The scope is specified by the number (K) of top-ranked documents presented to the users.

4.3 Compared Methods

We compare with a number of state-of-the-art methods: three popular methods (i.e., PLS [35], BLM [28], [29] and CCA [25], [30]) utilizing the pairwise information and four popular supervised methods (i.e., CDFE [60], GMLDA [28], GMMFA [28] and CCA-3V [31]) which take semantic category information into account. We also compare with the previous version of our method (Learning Coupled Feature Spaces, LCFS) [7].

PLS, BLM and CCA are three classical methods which use pairwise information to learn a common latent subspace across multi-modal data. In the common subspace, the similarity between different modalities of data can be measured. The above mentioned approaches enforce pair-wise closeness between different modalities of data in the common subspace.

CDFE, GMLDA, GMMFA and CCA-3V are supervised methods which exploit the label information. Due to the use of label information, CDFE, GMLDA, GMMFA and CCA-3V learn a discriminative common subspace. CCA-3V defines the similarity function between different modalities of data by setting different weights on the learnt representation to improve the performance. However, it does not perform feature selection on raw features of different modalities of data. In contrast, JFSSL resorts to ℓ_{21}-norm to perform feature selection. Although the ℓ_{21}-norm is minimized by an iteratively reweighted way, it is significantly different from the way of setting weights on representations. As a result, the sparse projections learned by JFSSL can lead to feature selection on raw features.

4.4 Parameter Setting

Note that the objective function in Eq. (4) mainly involves three parameters λ_1, λ_2 and β. λ_1 is the weighting parameter of the ℓ_{21}-norms, λ_2 is the weighting parameter of the

TABLE 1
MAP Comparison of Different Methods on the Pascal VOC Dataset

Methods	Image query	Text query	Average
PCA+PLS	0.2757	0.1997	0.2377
PCA+BLM	0.2667	0.2408	0.2538
PCA+CCA	0.2655	0.2215	0.2435
PCA+CDFE	0.2928	0.2211	0.2569
PCA+GMMFA	0.3090	0.2308	0.2699
GMMFA	0.2253	0.1695	0.1974
PCA+GMLDA	0.3094	0.2448	0.2771
PCA+CCA-3V	0.3146	0.2562	0.2854
LCFS	*0.3438*	*0.2674*	*0.3056*
JFSSL	**0.3607**	**0.2801**	**0.3204**

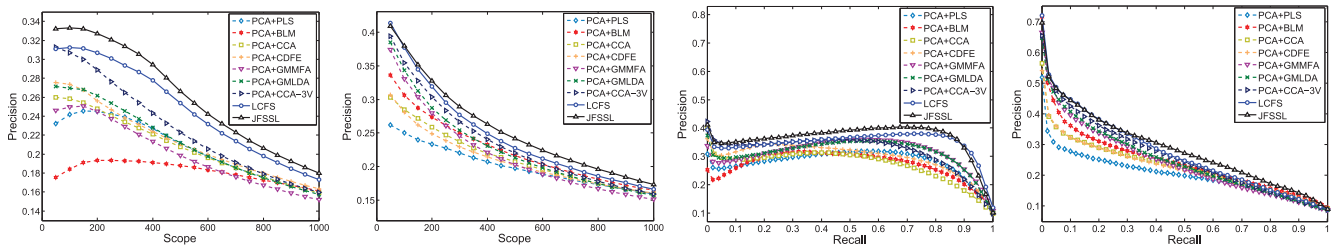
multimodal graph regularization, and β is the parameter which balances the effect of inter-modality and intra-modality similarity relationships. We tune them from $\{0.001, 0.01, 0.1, 1, 10, 100\}$ by cross validation. We will discuss the parameter sensitivity in the following. For the compared methods, we tune their parameters according to the corresponding literature.

4.5 Perfomance on Cross-Modal Retrieval

4.5.1 Results on Pascal VOC Dataset

As mentioned in Section 2, since the compared methods mainly focus on learning a common subspace, Principal Component Analysis (PCA) is performed on the original features to remove the redundancy in features. Our method can perform feature selection on different feature spaces simultaneously so that we do not perform PCA on the original features for our method and its previous version (LCFS). Table 1 shows the MAP scores achieved by PLS, BLM, CCA, CDFE, GMMFA, GMLDA, CCA-3V, LCFS and our method (JFSSL) on the Pascal VOC dataset. To illustrate the importance of PCA, the results of GMMFA without performing PCA on the original features are also reported in Table 1, which are much worse than those of performing PCA. We observe that our method (JFSSL) and its previous version (LCFS) outperform its several counterparts. This may be because JFSSL and LCFS select the relevant and discriminative features from different modalities simultaneously, while learning the common subspace. The learnt common subspace of JFSSL is more compact and effective by exploiting the inter-modality and intra-modality similarity relationships, which further improves the performance. From Table 1, we also see that CDFE, GMMFA, GMLDA, CCA-3V, LCFS and JFSSL perform better than PLS, BLM and CCA. This is because BLM, CCA and PLS only use pairwise information, but CDFE, GMMFA, GMLDA, CCA-3V, LCFS and JFSSL use class information, which provides much better separation between classes in the common subspace.

The corresponding precision-scope curves and precision-recall curves are plotted in Fig. 3. The scope (i.e., the top K

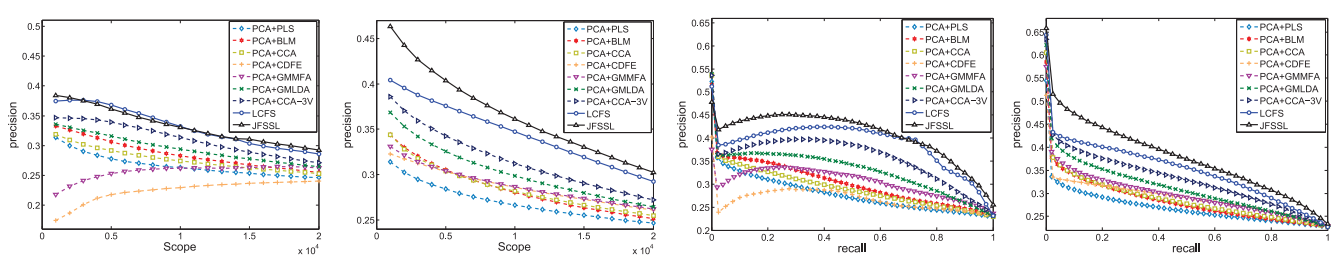


(a) Image query (b) Text query (c) Image query (d) Text query

Fig. 3. Performance of different methods on the Pascal VOC dataset, based on precision-scope curve (a-b) for K = 50 to 1,000 and precision-recall curve (c-d).

TABLE 2
MAP Comparison of Different Methods on the NUS-WIDE Dataset

Query	PCA+PLS	PCA+BLM	PCA+CCA	PCA+CDFE	PCA+GMMFA	PCA+GMLDA	PCA+CCA-3V	LCFS	JFSSL
Image	0.2752	0.2976	0.2872	0.2595	0.2983	0.3243	0.3513	*0.3830*	**0.4035**
Text	0.2661	0.2809	0.2840	0.2869	0.2939	0.3076	0.3260	*0.3460*	**0.3747**
Average	0.2706	0.2892	0.2856	0.2732	0.2961	0.3159	0.3386	*0.3645*	**0.3891**



(a) Image query (b) Text query (c) Image query (d) Text query

Fig. 4. Performance of different methods on the NUS-WIDE dataset, based on precision-scope curve (a-b) for K = 1,000 to 20,000 and precision-recall curve (c-d).

TABLE 3
MAP Comparison of Different Methods on the Wiki Dataset

Methods	Image query	Text query	Average
PLS	0.2402	0.1633	0.2032
BLM	0.2562	0.2023	0.2293
CCA	0.2549	0.1846	0.2198
CDFE	0.2655	0.2059	0.2357
GMMFA	0.2750	0.2139	0.2445
GMLDA	0.2751	0.2098	0.2425
CCA-3V	0.2752	0.2242	0.2497
LCFS	*0.2798*	*0.2141*	*0.2470*
JFSSL	**0.3063**	**0.2275**	**0.2669**

retrieved items) for the precision-scope curves varies from $K = 50$ to 1,000. Figs. 3a and 3b shows the performance of different methods based on the precision-scope curves for both forms of cross-modal retrieval tasks, i.e., Image query versus Text database and Text query vs. Image database. We observe that compared with its several counterparts, our method obtains better results for both tasks. Figs. 3c and 3d shows the performance of different methods based on the precision-recall curves, and our method also outperforms other methods for both forms of cross-modal retrieval. Fig. 6 shows the top nine images retrieved by JFSSL, CCA-3V, GMLDA and CCA respectively, given the tags "aeroplane+sky+building+shadow".

4.5.2 Results on NUS-WIDE Dataset

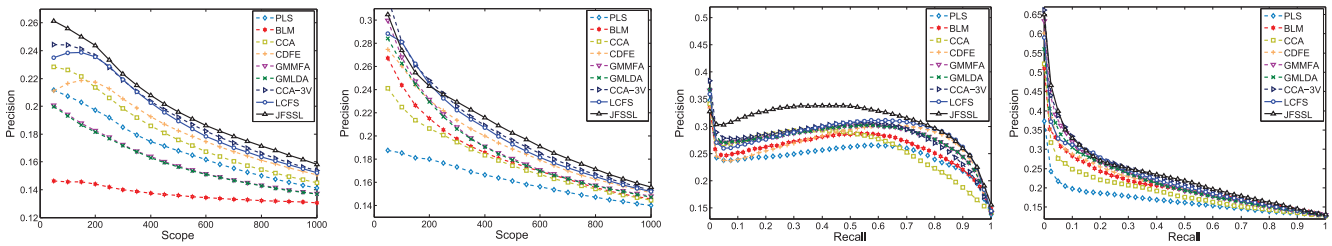
For the experiments on the NUS-WIDE dataset, Principal Component Analysis is also performed on the original features to remove redundancy in features for the compared methods. Table 2 shows the MAP scores achieved by PLS, BLM, CCA, CDFE, GMMFA, GMLDA, CCA-3V, LCFS and JFSSL on the NUS-WIDE dataset. We observe that our method outperforms its several counterparts. The experimental results are similar to those on the Pascal VOC dataset.

Fig. 4 shows the corresponding precision-scope curves (a-b) and precision-recall curves (c-d) for both forms of cross-modal retrieval tasks, i.e., Image query versus Text database and Text query versus Image database. We observe that compared with its several counterparts, our method performs best for both tasks.

4.5.3 Results on Wiki Dataset

Due to the low dimensions of image and text features themselves on the Wiki dataset, PCA is not used here to reduce the dimensions of the original features here. Table 3 shows the MAP scores of different approaches on the Wiki dataset. LCFS achieves MAP scores of 0.2798 and 0.2141 for the image query and text query respectively, only a little bit better than those of GMMFA and GMLDA. The reason is that the dimensions of image and text features are low so that the ℓ_{21}-norms of our method for feature selection hardly takes effect. JFSSL achieves better performance (MAP scores of 0.3063 and 0.2275 for the image and text query respectively) by exploiting the inter-modality and intra-modality similarity relationships through the multimodal graph regularization. We also see that CDFE, GMMFA, GMLDA, CCA-3V and our methods (LCFS and JFSSL) perform better than PLS, BLM, and CCA because of taking the semantic information into account.

The corresponding precision-scope curves and precision-recall curves are plotted in Fig. 5. From the precision-scope curves in (a-b), we can see that for both forms of



(a) Image query (b) Text query (c) Image query (d) Text query

Fig. 5. Performance of different methods on the Wiki dataset, based on precision-scope curve (a-b) for K = 50 to 1,000 and precision-recall curve (c-d).

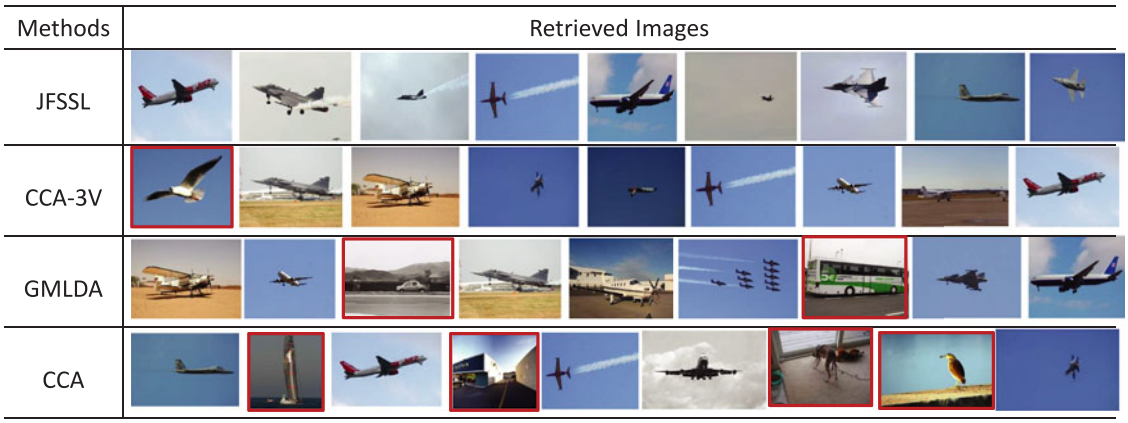


Fig. 6. An example of cross-modal retrieval using text query (i.e., the tags "aeroplane+sky+building+shadow") on the Pascal VOC dataset. Red border indicates a incorrect retrieval result.

Fig. 7. Four examples of cross-modal retrieval using text queries on the Wiki dataset. The text query and ground truth image are shown on the left; the top three images retrieved by our method are presented at the right.

cross-modal retrieval, our method finds more number of correct matches in the top K documents than its compared methods. Our method also obtains better performance in terms of precision-recall curve, as shown in Figs. 5c and 5d.

Fig. 7 shows four examples of text queries and the top three images retrieved by our method. In each case, the query text and its paired image are shown at the left, and the top three images are shown at the right. Note that our method finds the closest matches at semantic level, i.e., the common subspace pre-computed by class labels. The retrieved images are perceived as belonging to the same category of the query text ("Music" at the top row, "Warfare" at the second row, "Biology" at the third row, and "Literature & theatre" at the bottom row).

From the experiments on the three datasets, we can draw the following conclusions:

- CDFE, GMMFA, GMLDA, CCA-3V, JFSSL and its previous version achieve better results than PLS, BLM and CCA. The reason is that PLS, BLM and CCA only care about pair-wise closeness in the common subspace, but CDFE, GMMFA, GMLDA, CCA-3V, JFSSL and its previous version utilize class information to obtain much better separation between classes in the common subspace, which improves the performance. Hence, it is helpful for cross-modal retrieval to learn a discriminative common subspace.
- Compared with several state-of-the-art methods, our proposed method (JFSSL) consistently performs better. One of the reasons is that JFSSL selects the relevant and discriminative features from different modalities simultaneously, while learning a common subspace. It is important for cross-modal retrieval to select the relevant and discriminative features because redundant and irrelevant features affect the performance much. The multimodal graph regularization further improves the performance of the proposed method, suggesting the usefulness of inter-modality and intra-modality similarity relationships among different modalities of data objects.

TABLE 5
MAP Comparison of Different Methods on the Wiki Dataset in the Three-Modality Case

Query	Modality A	Modality B	Modality C
PLS	0.1629	0.1653	0.2412
BLM	0.1673	0.2167	0.2607
CCA	0.1733	0.1722	0.2434
CDFE	0.1882	0.1836	0.2548
GMMFA	0.2005	0.1961	0.2551
GMLDA	0.1841	0.1700	0.2525
CCA-3V	0.2301	0.1720	0.2665
LCFS	*0.2292*	*0.3065*	*0.3072*
JFSSL	**0.2636**	**0.3203**	**0.3354**

4.6 Image-to-Image Retrieval

Although our main goal is cross-modal retrieval, we also perform the image-to-image retrieval task on the Wiki dataset. Given a query image, we project its raw feature into the learnt space, and use it to retrieve the most similar images from the database. We randomly select 200 images form the testing set as the query, and take the rest of images as the database.

Table 4 shows the MAP comparison of different methods on the Wiki dataset for the image-to-image retrieval. The subspace learning methods perform better than raw features. It indicates that there is a benefit to explicitly map the raw feature into the learnt space. It may be because the texts carry the information which could be complementary to images. As a result, we could obtain more accurate representations in the learnt space than raw features. Moreover, JFSSL performs best. The reason is that it not only performs feature selection and subspace learning jointly, but also incorporates the similarity relationships into the learnt representation.

4.7 Three-Modality Case

Since the proposed framework is formulated for more than two modalities, we evaluate its effectiveness in the three-modality case on the Wiki dataset. Currently to the best of our knowledge, there are no three or more modalities of datasets available publicly in the literature. The Wiki dataset contains two modalities of data: text and image. To simulate a three-modality setting, 4,096-dimensional CNN features of images are extracted by Caffe [61] as another virtual modality [62]. Here 128-dim SIFT histogram, 10-dim LDA feature and 4,096-dimensional CNN features are used as Modality A, Modality B and Modality C respectively. Since the compared methods cannot handle the three-modality case directly, we apply them to the three-modality case in the following way: take $A \rightarrow (B, C)$ as example, A is served as the query and retrieved results contain data from

TABLE 4
MAP Comparison of Different Methods on the Wiki Dataset for the Image-to-Image Retrieval Task

Task	Raw Feature	PLS	BLM	CCA	CDFE	GMMFA	GMLDA	CCA-3V	LCFS	JFSSL
Image To Image	0.1504	0.1592	0.1565	0.1594	0.1571	0.1613	0.1615	0.1657	*0.1678*	**0.1734**

TABLE 6
Evaluation of Regularization Terms on the Wiki Dataset in the Three-Modality Case

Query	Modality A	Modality B	Modality C
JFSSL	0.2636	0.3203	0.3354
JFSSL(with ℓ_2 norm)	0.2486	0.3097	0.3294
JFSSL($\lambda_2 = 0$)	0.2347	0.3085	0.3040
JFSSL($\beta = 0$)	0.2513	0.3166	0.3193

B and C. For A and B, we use the projections learned from A and B to map A and B into the learnt space, in which the presentations are Z_A and Z_B respectively. Then, we learn projections from A and C, and project C into the learnt space, in which the presentation of C is Z_C. Through exploiting the learnt representations, we can use Z_A as the query to retrieve Z_B and Z_C.

Table 5 shows the MAP comparison on the Wiki dataset in the three-modality case. We can see that our JFSSL outperforms the compared methods in three cross-modal retrieval tasks. This is mainly due to the fact that our formulation can model the correlations between different modalities more accurately in the three-modality case. The compared methods are designed for only the two-modality case. Although CCA-3V is a 3-view method, it treats supervised information (i.e., class label) as a view. Hence, CCA-3V is essentially designed for the two-modality case. These compared methods are not suitable for the three-modality case, which leads to poor performance accordingly.

4.8 Evaluation of Regularization Terms

We evaluate the importance of the ℓ_{21}-norm term and the multimodal graph regularization term in our framework. Table 6 shows the MAP comparison on the Wiki dataset in the three-modality case when replacing or removing the regularization terms. We test the cross-modal retrieval performance when the ℓ_{21}-norm is replaced with ℓ_2 norm. We can see that the ℓ_{21}-norm is indeed useful, which can learn sparse projection matrix for different modalities of data simultaneously. We also evaluate the framework without the multimodal graph regularization term (and without the intra-modal similarity relationships). It can be seen that the multimodal graph regularization term, which exploits the inter-modality and intra-modality similarity relationships, is beneficial for the cross-modal retrieval.

4.9 Performance with Different Types of Features

We also test the cross-modal retrieval performance with different types of features for images and texts on the Wiki dataset. Besides the features provided by the Wiki dataset itself, 4,096-dimensional CNN features for images are extracted by Caffe, and 5,000-dimensional feature vectors for texts are extracted by using the bag of words representation with the TF-IDF weighting scheme. Table 7 shows the MAP scores of GMLDA, CCA-3V, LCFS and JFSSL with different types of features on the Wiki dataset. PCA is performed on CNN and TF-IDF features for GMLDA and CCA-3V. It can be seen that all methods achieve better results when using the CNN features. This is because CNN features are more powerful for image representation, which has been proved in many fields. As expected, our proposed JFSSL still outperforms its two major competitors (i.e., GMLDA and CCA-3V).

TABLE 7
MAP Comparison with Different Features on the Wiki Dataset

Query	Methods	Features (Image/Text)			
		SIFT/ LDA	CNN/ LDA	SIFT/ TF-IDF	CNN/ TF-IDF
Image	GMLDA	0.2751	0.4084	0.2782	0.4455
	CCA-3V	0.2752	0.4049	0.2862	0.4370
	LCFS	*0.2798*	*0.4132*	*0.2978*	*0.4553*
	JFSSL	**0.3063**	**0.4279**	**0.3080**	**0.4670**
Text	GMLDA	0.2098	0.3693	0.1925	0.3661
	CCA-3V	0.2242	0.3651	0.2238	0.3832
	LCFS	*0.2141*	*0.3845*	*0.2134*	*0.3978*
	JFSSL	**0.2275**	**0.3957**	**0.2257**	**0.4102**

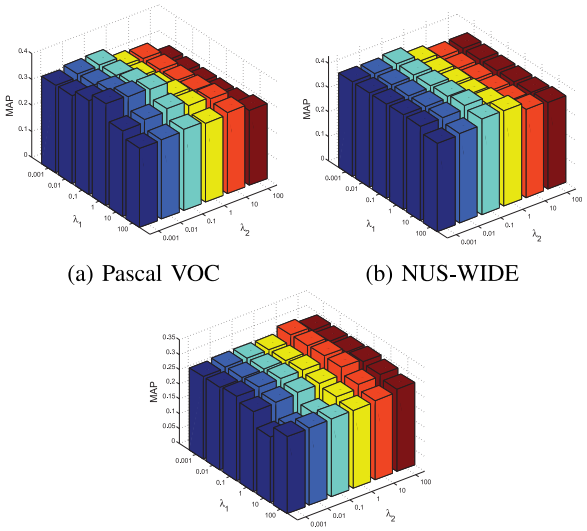
Now we look into the selected features by JFSSL for the Wiki dataset. Since the image features are difficult to illustrate, we only demonstrate the selected features by textual words when using the TF-IDF features. Note that for the learnt projection $\mathbf{U} \in \mathbb{R}^{d \times c}$, each column in $\mathbf{U}$ weights the contribution of all words to the corresponding semantics. The corresponding words with large value $\mathbf{U}_{ij}$ are the selected ones. We sort each column by which the selected words are ranked ahead. We present the selected words on some categories from the Wiki dataset in Fig. 8. It can be seen that the selected words are relevant to the corresponding semantic concepts.

4.10 Parameter Sensitivity Analysis

There are several parameters involved in the proposed method, i.e., the weighting parameters λ_1, λ_2 and the trade-off parameter β. We tune these three parameters in the range of $\{0.001, 0.01, 0.1, 1, 10, 100\}$. First, we fix β and report the performance when λ_1 and λ_2 are changing. The experimental results on the three datasets for Image query versus Text database and Text query versus Image database tasks are shown in Figs. 9 and 10, respectively. We observe that the performance of our algorithm varies when the parameters

Category	Selected Words
Art	ARCHITECTURE PAINTING PHOTOGRAPHS IMAGES MEMORIAL SWEDISH PORTRAIT PAINTED CLASSIC TOWER HUDSON
Biology	BONES GENUS PREDATORS EXTINCTION HABITATS BREEDING SEED POPULATIONS POTENTIAL DARWIN ZOO BREED
Literature	READERS NOVELS COMIC DESIRE PRINTED PATENT POETRY MANUSCRIPT PLOT JURY POET EXPLAINS EDGAR PUBLISHER
Media	FILMS BROADWAY FILMING ACTRESS DVD CAR DOCTOR NEWSPAPER CHARLIE MUTUAL SIMPSON REFUGEES SERIAL JACK
Sport	PLAYER BASEBALL CHAMPIONSHIP PLAYERS MANAGER PLAYOFFS ESPN AMATEUR RACE CRICKET INNINGS MOVES HUNTER
Warfare	PASSENGERS INFANTRY MEMOIRS ALLIES BATTLESHIPS DELIVERED RAAF 1915 MISSION TANK CASUALTIES COMBAT TASK

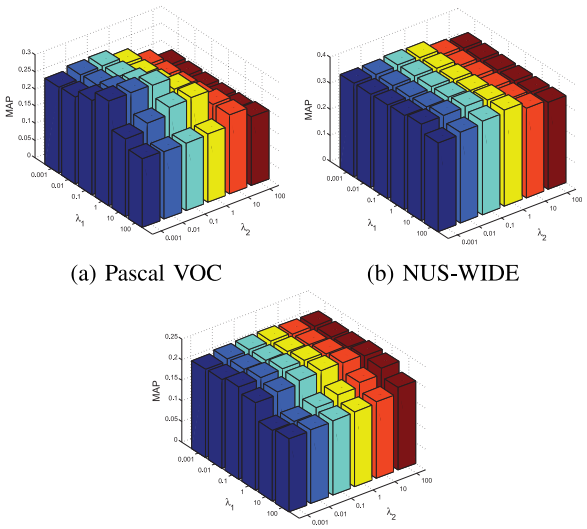
Fig. 8. Selected words on some categories by JFSSL from the Wiki dataset.



(a) Pascal VOC (b) NUS-WIDE

(c) Wiki

Fig. 9. Performance variation for the Image query versus Text database task with respect to λ_1 and λ_2 when we fix β for the Pascal VOC, NUS-WIDE and Wiki datasets respectively.



(a) Pascal VOC (b) NUS-WIDE

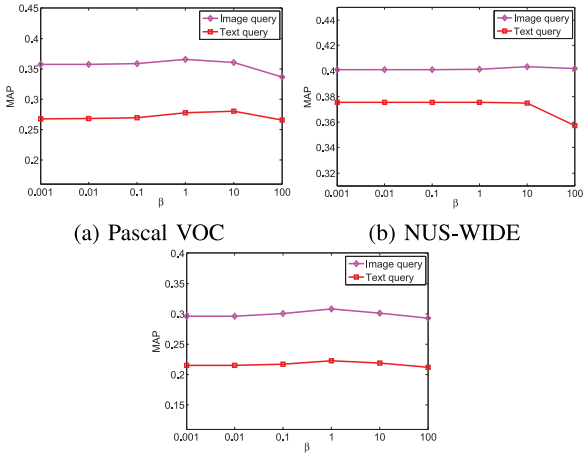
(c) Wiki

Fig. 10. Performance variation for the Text query versus Image database task with respect to λ_1 and λ_2 when we fix β for the Pascal VOC, NUS-WIDE and Wiki datasets respectively.

are different. Generally speaking, the proposed method obtains better performance when λ_1 is 0.1 to 10 for the datasets. This is because λ_1 controls the sparsity of the projection matrices, a too small value cannot select out the relevant and discriminative features, but a too large value will lead to remove some useful features. We also observe similar results for the parameter λ_2, which is the weighting parameter of the multimodal graph regularization. We also fix λ_1 and λ_2 to test the performance on the three datasets for both forms of tasks when β is changing. As shown in Fig. 11, the proposed method is not much sensitive to the trade-off parameter β.

4.11 Convergence

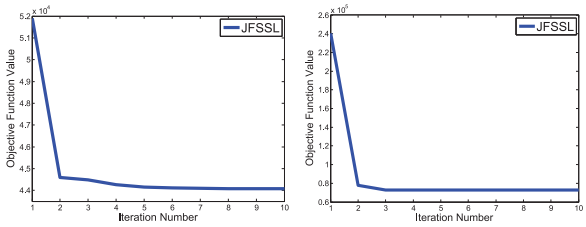
An iterative optimization algorithm is proposed to solve the objective function. For practical applications, it is interesting



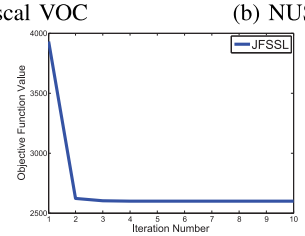
(a) Pascal VOC (b) NUS-WIDE

(c) Wiki

Fig. 11. Performance variation with respect to β when we fix λ_1 and λ_2 for the Pascal VOC, NUS-WIDE and Wiki datasets respectively.



(a) Pascal VOC (b) NUS-WIDE



(c) Wiki

Fig. 12. Convergence curves of the objective function value in Eq. (6) using Algorithm 1. The figure shows that the objective function value monotonically decreases until convergence by applying the iterative algorithm.

to see how fast the iterative algorithm converges. In Fig. 12, we plot the convergence curves of our iterative algorithm with respect to the objective function value of (4) at each iteration on Pascal VOC, NUS-WIDE, and Wiki datasets, respectively. In this figure, the objective function value of (4) monotonically decreases at each iteration. More specifically, the algorithm generally converges within about five iterations for all datasets. The running time[2] of convergence on the Pascal VOC, NUS-WIDE, and Wiki datasets are 12.8 seconds, 0.6 hour, and 1.7 seconds, respectively. The bottleneck lies in the multimodal graph construction, so our future work is to further reduce the running time of graph construction.

5 Conclusion

In this paper, we have proposed a novel joint learning framework to solve the problem of cross-modal retrieval, which consists of subspace learning for different modalities, the ℓ_{21}-norms for coupled feature selection, and the

2. We run our matlab code on a 2-core Xeon 2.67 GHz workstation with 128 GB RAM.

multimodal graph regularization for preserving the inter-modality and intra-modality similarity. Under the proposed framework, different projection matrices are learnt to map different modal data into a common subspace, and relevant and discriminative features for different spaces are selected simultaneously in the projection procedure. Furthermore, the inter-modality and intra-modality similarity are well preserved through the multimodal graph regularization while mapping. To solve this joint learning problem, we have presented an iterative optimization algorithm along with its convergence analysis. Experimental results on three cross-modal datasets have demonstrated that the proposed method performs better than several relevant state-of-the-art subspace methods.

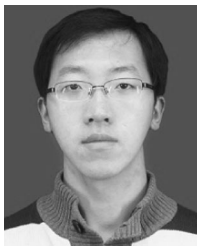
ACKNOWLEDGMENTS

This work is jointly supported by National Basic Research Program of China (2012CB316305), and National Natural Science Foundation of China (61420106015,61525306). Liang Wang is the corresponding author of this paper.

REFERENCES

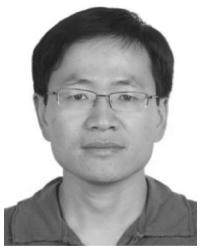
[1] R. Bekkerman and J. Jeon, "Multi-modal clustering for multimedia collection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[2] P. K. Atrey and M. A. Hossain, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010.
[3] Y. F. Fu, T. M. Hospedales, T. Xiang, and S. G. Gong, "Learning multimodal latent attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 303–316, Feb. 2014.
[4] C. Xu, D. C. Tao, and C. Xu, "Large-margin multi-view information bottleneck," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1559–1572, Aug. 2014.
[5] F. Wu, Y. Yuan, X. Liu, J. Shao, Y. Zhuang, and Z. Zhang, "The heterogeneous feature selection with structual sparsity for multimedia annotation and hashing: A survey," *Int. J. Multimedia Inf. Retrieval*, vol. 1, no. 1, pp. 3–15, 2012.
[6] F. Wu, Y. Yuan, and Y. Zhuang, "Heterogeneous feature selection by group lasso with logistic regression," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 983–986.
[7] K. Y. Wang, R. He, W. Wang, L. Wang, and T. N. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2088–2095.
[8] L. Sun, S. Ji, and J. Ye, "A least squares formulation for canonical correlation analysis," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1024–1031.
[9] R. He, T. N. Tan, L. Wang, and W. Zheng, "ℓ_{21} regularized correntropy for robust feature selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2504–2511.
[10] D. Blei and M. Jordan, "Modeling annotated data," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 127–134.
[11] D. Putthividhy, H. Attias, and S. Nagarajan, "Topic regression multi-modal latent Dirichlet allocation for image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3408–3415.
[12] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[13] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2407–2414.
[14] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2231–2239.
[15] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 604–611.
[16] W. Wu, J. Xu, and H. Li, "Learning similarity function between objects in heterogeneous spaces," Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-2010-86, 2010.
[17] A. Mignon and F. Jurie, "CMML: A new metric learning approach for cross modal matching," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 1–14.
[18] N. Quadrianto and C. H. Lampert, "Learning multi-view neighborhood preserving projections," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 425–432.
[19] X. H. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. AAAI*, 2013, pp. 1198–1204.
[20] X. Y. Lu, F. Wu, S. L. Tang, Z. F. Zhang, X. F. He, and Y. T. Zhuang, "A low rank structural large margin method for cross-modal ranking," in *Proc. 36th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 433–442.
[21] F. Wu, X. Y. Lu, Z. F. Zhang, S. C. Yan, Y. Rui, and Y. T. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 877–886.
[22] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using Similarity-sensitive hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3594–3601.
[23] Y. Zhen and D.-Y. Yeung, "Co-regularized hashing for multimodal data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1385–1393.
[24] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.
[25] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Cannonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[26] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.
[27] R. Rosipal and N. Kramer, "Overview and recent advances in partial least squares," in *Proc. Statistical Optimization Perspectives Workshop: Subspace, Latent Struct. Feature Selection*, 2006, pp. 34–51.
[28] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2160–2167.
[29] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, 2000.
[30] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
[31] Y. C. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.
[32] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.
[33] R. Udupa and M. Khapra, "Improving the multilingual user experience of wikipedia using cross-language name search," in *Proc. Human Language Technol.: Annu. Conf. North Am. Chapter Assoc. Comput. Linguistics*, 2010, pp. 492–500.
[34] A. Li, S. Shan, X. Chen, and W. Gao, "Face recognition based on non-corresponding region matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1060–1067.
[35] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 593–600.
[36] Y. Chen, L. Wang, W. Wang, and Z. Zhang, "Continuum regression for cross-modal multimedia retrieval," in *Proc. IEEE Int. Conf. Image Process.*, 2012, pp. 1949–1952.
[37] V. Mahadevan, C.-W. Wong, T. T. Liu, N. Vasconcelos, and L. K. Saul, "Maximum covariance unfolding: Manifold learning for bimodal data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 918–926.
[38] X. B. Mao, B. B. Lin, D. Cai, X. F. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 897–906.
[39] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. 9th Eur. Conf. Comput. Vis.*, pp. 13–26, 2006.

[40] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI*, 2013, pp. 1070–1076.
[41] Y. Yang, D. Xu, F. Nie, J. B. Luo, and Y. T. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 175–184.
[42] Y. T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 221–229, Feb. 2008.
[43] Z. W. Lu and Y. X. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 306–325, 2013.
[44] Z. W. Lu and Y. Peng, "Unified constraint propagation on multi-view data," in *Proc. AAAI*, 2013, pp. 640–646.
[45] X. H. Zhai, X. Y. Peng, and J. G. 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.
[46] S. Bengio, J. Weston, and D. Grangier, "Label embedding trees for large multi-class tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 163–171.
[47] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
[48] J. Weston, S. Bengio, and N. Usunier, "WSABIE: Scaling up to large vocabulary image annotation," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 2764–2770.
[49] K. Q. Weinberger and O. Chapelle, "Large margin taxonomy embedding for document categorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1737–1744.
[50] A. Frome, G. S. Corrado, J. Shlens, et al., "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
[51] D. Cai, X. F. He, and J. W. Han, "Spectral regression for efficient regularized subspace learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[52] R. He, W. Zheng, and B. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.
[53] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Sci. Comput.*, vol. 27, no. 3, pp. 937–966, 2005.
[54] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint ℓ_{21}-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[55] S. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1145–1158, Jun. 2012.
[56] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.
[57] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. B. Luo, and Y. T. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 395–404.
[58] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[59] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.
[60] D. H. Lin and X. O. Tang, "Inter-modality face recognition," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 13–26.
[61] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[62] G. G. Ding, Y. C. Guo, and J. L. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 4321–4328.

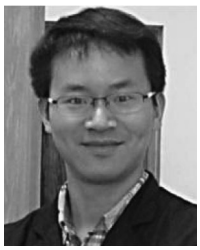


Kaiye Wang received the BS degree in computer science and technology from Jilin University, the MS degree in computer application technology from Jilin University, and PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences in 2009, 2012, and 2015, respectively. His research interests include cross-modal retrieval/hashing, multiview learning, and representation learning.

Ran He received the BE degree in computer science from Dalian University of Technology, the MS degree in computer science from Dalian University of Technology, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2001, 2004, and 2009, respectively. Since September 2010, he has been joined NLPR, where he is currently a project professor. He has widely published at highly ranked international journals, such as *TPAMI, TKDE, TIP*, CVPR and AAAI. He is a senior member of the IEEE.



Liang Wang received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CAS) in 2004. He is currently a full Professor of Hundred Talents Program at the NLPR, Institute of Automation, CAS, China. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published at highly ranked international journals such as *IEEE TPAMI* and *IEEE TIP*, and leading international conferences such as CVPR, ICCV, and ICDM. He is a senior member of the IEEE.



Wei Wang received the BE degree from the Department of Automation, Wuhan University in 2005, and the PhD degree from the School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences (GUCAS) in 2011. Since July 2011, he has been joined NLPR, where he is currently an assistant professor. His research interests focus on computer vision, pattern recognition, visual attention, and deep learning.

Tieniu Tan received the BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the MSc and PhD degrees in electronic engineering from Imperial College London, United Kingdom, in 1986 and 1989, respectively. He was the director general in the CAS Institute of Automation from 2000 to 2007, where he is currently a professor and the former director (1998-2013) in the NLPR since 1998. He also serves as a vice president in the CAS. He has published 11 edited books or monographs and more than 400 research papers in refereed international journals and conferences in the areas of image processing, computer vision and pattern recognition. He is a fellow of the CAS, TWAS, IEEE, and IAPR, and an International Fellow of the United Kingdom Royal Academy of Engineering.