Cross-Modal Retrieval Using Multiordered Discriminative Structured Subspace Learning

Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian

***Abstract*—This paper proposes a novel method for cross-modal retrieval. In addition to the traditional vector (text)-to-vector (image) framework, we adopt a matrix (text)-to-matrix (image) framework to faithfully characterize the structures of different feature spaces. Moreover, we propose a novel metric learning framework to learn a discriminative structured subspace, in which the underlying data distribution is preserved for ensuring a desirable metric. Concretely, there are three steps for the proposed method. First, the multiorder statistics are used to represent images and texts for enriching the feature information. We jointly use the covariance (second-order), mean (first-order), and bags of visual (textual) features (zeroth-order) to characterize each image and text. Second, considering that the heterogeneous covariance matrices lie on the different *Riemannian* manifolds and the other features on the different *Euclidean* spaces, respectively, we propose a unified metric learning framework integrating multiple distance metrics, one for each order statistical feature. This framework preserves the underlying data distribution and exploits complementary information for better matching heterogeneous data. Finally, the similarity between the different modalities can be measured by transforming the multiorder statistical features to the common subspace. The performance of the proposed method over the previous methods has been demonstrated through the experiments on two public datasets.**

***Index Terms*—Cross-modal retrieval, documents and images, multimedia.**

Manuscript received May 27, 2016; revised October 2, 2016; accepted December 12, 2016. Date of publication December 29, 2016; date of current version May 13, 2017. This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2015CB351800 and Grant 2012CB316400, in part by the National Natural Science Foundation of China under Grant 61572465, Grant 61332016, Grant 61429201, Grant 61620106009, Grant U1636214, and Grant 61303153, and in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under Grant QYZDJ-SSW-SYS013. The work of Q. Tian was supported in part by the ARO Grant W911NF-15-1-0290 and the Faculty Research Gift Awards by the NEC Laboratories of America and Blippar. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Enrico Magli. *(Corresponding author: Bingpeng Ma.)*

L. Zhang, B. Ma, G. Li, and Q. Huang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China, and and also with the Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhangliang14@mails.ucas.ac.cn; bpma@ucas.ac.cn; liguorong@ucas.ac.cn; qmhuang@ucas.ac.cn).

Q. Tian is with the Department of Computer Sciences, University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2646219

I. Introduction

As THE major component of big data, multi-modal data including image, text, video and audio have emerged on the Internet rapidly, and it is now imperative to exploit the correlations among multimedia data. Consequently, cross-modal retrieval has attracted considerable attention in recent years. In the multimedia research field, the cross-modal learning has many practical applications, such as image retrieval [1], [2], image annotation [3] and multi-modal video retrieval [4]. Specifically, we focus on exploiting the correlation between image modality and text modality, which has been actively studied in many works [5]–[6].

As is well known, the different multimedia data reside in different feature spaces. Hence, the key problem for cross-modal retrieval is how to model the correlations among the multimodal data. A large number of methods have been proposed to alleviate this problem by learning a common subspace for the multimedia data. They represent the multimedia data as high-dimensional vectors, and exploit the correlations among the multimedia data by learning the optimal transformations for these vectors. Generally, these methods can be mainly classified into four kinds of directions [7]. First, some methods learn the maximal correlations among different modalities to obtain the common subspace [1], [5], [8]–[12]. The most popular method could be canonical correlation analysis (CCA) [1]. The motivation of CCA is to learn a common subspace which maximizes the correlations between the projected vectors of two modalities. Second, the manifold learning methods are also adopted to obtain the common subspace [13]–[14]. Since the high dimensional data may embed in a lower dimensional intrinsic space, the manifold learning methods project the different modalities into a common manifold by learning their underlying manifold representation. Third, different from the above methods, the third direction learns the common subspace by using the technique of learning to rank [6], [15]–[18]. These methods maximize a criterion related to the ultimate retrieval performance and obtain the common subspace by large margin learning with certain ranking criteria. Finally, the semantic content implied in the multimedia data can be refined as class labels, which directly reveal the semantic information of multimedia data. Considering this, many works learn the discriminative subspace by using the valuable class information [19]–[22].

The above methods cannot be directly applied to higher order features such as two-dimensional matrices, and would

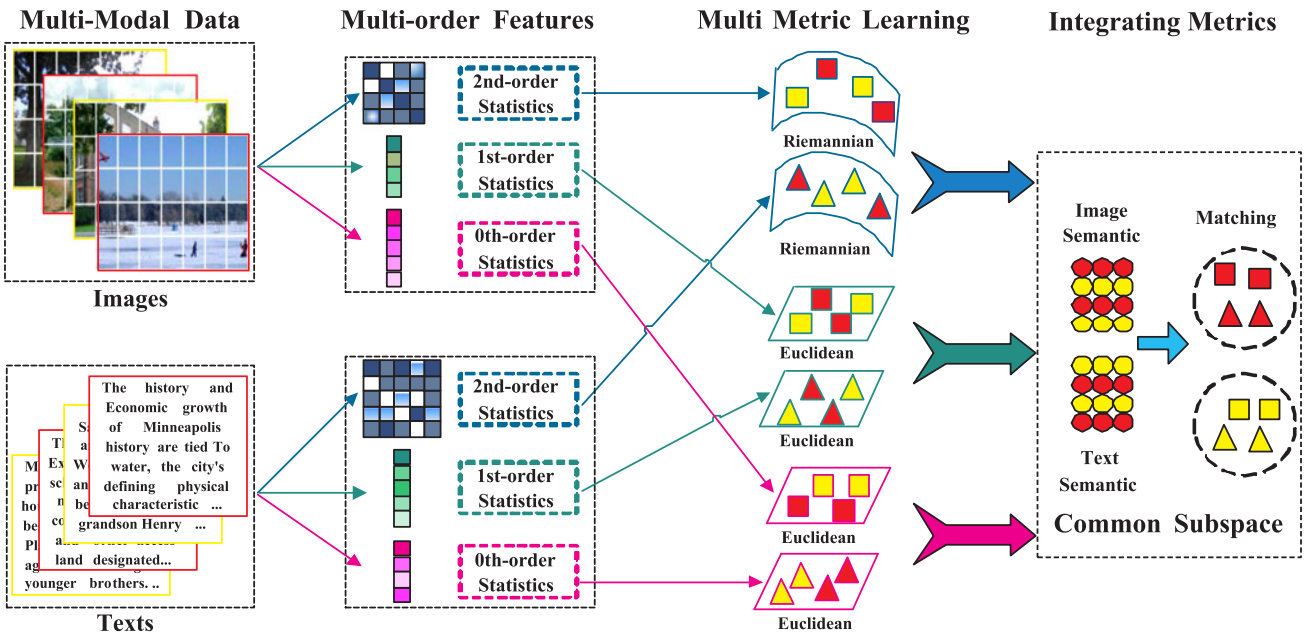


Fig. 1. Algorithmic flowchart of the proposed method. For the sake of illustrative simplicity, we show only two samples for each class in different modalities. Shapes represent modalities (i.e., text and image), the same color indicates relevant semantics. The multi-order statistical features are first adopted to represent each image and each text, then multilocal metrics (*Riemannian*-to-*Riemannian* and *Euclidean*-to-*Euclidean*) are conducted to exploit the semantic relationship for each statistical features. Furthermore, ensemble learning is applied to integrate local metrics and is exploited the complementary information of different statistical features. Finally, the semantic features are learned on a low-dimensional latent semantic space, in which two samples with similar semantics are close to each other.

require some pre-processing (e.g., vectorization of matrices). In fact, the higher order representation (e.g., covariance matrices) can characterize the structure information of feature space because it faithfully exploits the correlations among the entries of feature vector. The popular feature extraction methods often represent each multimedia data as a set of local feature descriptors, e.g., SIFT algorithm for visual modality and *word2vec* model for textual modality. Since each local feature descriptor is obtained by synthesizing various information contained in a data patch, we consider that each local feature descriptor encodes the local spatial information of a data patch. Therefore, it is important to exploit the structure informations of the different feature spaces, which are usually ignored in the process of vectorization for matrices.

To exploit the structure information of different feature spaces and capture the discriminative semantic information, we propose a novel metric learning framework named Multi-ordered Discriminative Structured Subspace Learning (MDSSL) to enhance the correlations among the multi-modal data. In MDSSL, we represent each multimedia data as a set of local feature descriptors by adopting the dense SIFT algorithm [23] and *word2vec* model [24]. Then we compute the holistic multi-order statistics as the features of the multimedia data. The 2nd-order statistical features faithfully characterize the structure information of the different feature spaces such that they can encode the feature correlations specific to each class. Furthermore, the 0th-order and 1st-order statistics are also computed for enriching the feature information because the different order statistical features characterize the feature space from the different aspects. Hence, the multi-order statistical features can extract rich information from low-level features to high-level semantic features such that the correlation between the different modalities can be further strengthened.

Since the multi-order statistical features lie on the different *Euclidean* spaces and *Riemannian* manifolds, we propose a novel metric learning framework to learn a discriminative subspace for heterogeneous data. This framework can preserve the underlying data distributions in the learning stage, and this strategy helps to learn an optimal metric for heterogeneous data. Furthermore, it can effectively integrate multiple distance metrics and exploit complementary information to better match the heterogeneous data for cross-modal retrieval. Fig. 1 shows the flowchart of the proposed method.

In summary, with the attractive and distinct advantages of MDSSL, the main contributions of this paper are:

1) A novel matrix-to-matrix framework is proposed for the cross-modal problem, which is formulated as matching points from heterogeneous *Riemannian* manifolds. By this way, the structure informations of heterogeneous feature spaces are effectively exploited simultaneously.
2) The multi-order statistical features are computed for the different modalities. Since different order statistics characterize features space from the different perspectives, integrating these features provides complementary information which is helpful to discriminate samples from different classes.
3) A discriminative structured subspace learning is proposed to make better use of the information from the multi-order statistical features. This framework not only can match heterogeneous data with only one-order statistical features, but also can work well in multi-order statistical features.

II. Related Work

The main problem of cross-modal retrieval is to exploit the semantic relationship among the different modalities. Since texts and images are the different modalities, they cannot be matched directly with each other. Thus, most of the recent studies are concentrated on learning the common subspace.

One popular approach is to obtain the common subspace by maximizing correlations between different modalities. Especially, CCA [1] seeks a pair of linear transformations to maximize the correlations between two modalities. Based on the good performance of CCA, many extensions and applications have been developed. For example, a logistic regression is adopted to compute the posterior probabilities assigned to all semantic categories after obtaining the maximally correlated subspace between two modalities [5]. Generalized multiview analysis (GMA) [10] is the supervised extension of CCA. GMA solves a joint, relaxed quadratic constrained program over the different feature spaces to obtain a single (non-) linear subspace. Its extensions, GMLDA and GMMFA, have been shown very good performance on cross-media retrieval.

An alternative relies on the manifold learning methods to obtain the common subspace. Based on the manifold alignment, the method of [13] constructs the transformations to link the different feature spaces in order to transfer knowledge across domains. In [25], the authors learn a common low dimensional embedding which can maximize the multi-modal correlations and preserve the local distances simultaneously. Additionally, parallel filed alignment [26] is proposed to project the correlation of heterogeneous media data into intermediate latent semantic spaces and preserve the metric of data manifolds during the process of manifold alignment.

Different from the above methods, the common subspace can also be obtained by learning to rank. Grangier *et al.* propose a method named passive-aggressive model for image retrieval (PAMIR) [17], which is the first attempt to address the problem of ranking images by text queries directly. PAMIR formulates the cross-modal retrieval problem similar as RankSVM [27] and derives an efficient training procedure by adapting the Passive-Aggressive algorithm. In [16], Supervised semantic indexing (SSI) defines a set of linear low-rank models to exploit the correlations between words. However, PAMIR and SSI are unidirectional ranking based methods. These methods only capture the correlation between two modalities from one direction of retrieval, and their generalization performance is limited since they do not capture the latent structure of the query modality. Considering this, Wu *et al.* [15] propose a bi-directional cross-media semantic representation model (Bi-CMSRM). which employs the structural SVM [28] to support the optimization of various ranking evaluation measures under a unified framework. Bi-CMSRM obtains a latent space embedding by learning the structural large margin to optimize the bi-directional listwise ranking loss simultaneously.

Besides, considering that the class labels can help to model the correlations between two modalities, Kang *et al.* [19] use the valuable class information to learn the consistent feature representation and discover useful information from the unpaired samples. Wang *et al.* [20] learn the coupled feature spaces (LCFS) for two modalities by optimizing a half quadratic labeling error.

It's worth mentioning that many deep models [29]–[30] have been proposed to model the correlation between the multimedia data. Specifically, Deep CCA [29] adopts the deep networks to learn flexible nonlinear representations for the multi-modal data. Deep boltzmann machines [31] is proposed to exploit the correlation by learning deep-based shared representation.

In summary, most of the existing methods learn the common subspace by using vector-to-vector framework. They are limited in characterizing the structure information of feature spaces of multimedia data. Compared with deep visual features [32], which characterize the local structure information by increasing the number of layers or revising the pooling layers, MDSSL simply uses the 2nd-order statistical features to effectively exploit discriminant information. Nevertheless, MDSSL can achieve comparable retrieval performances when we utilize the visual features extracted by deep models, e.g., convolutional neural network (CNN) [32]. Furthermore, MDSSL can learn the common subspace for multi-order statistical features. On one hand, the multi-order statistical features can characterize the feature spaces more faithfully such that objects belonging to different classes can be better discriminated. On the other hand, by simultaneously integrating multiple metrics, the complementary information can be exploited to better characterize two modalities for cross-modal retrieval.

III. Sample Modeling With Multiorder Statistics

In this section, we first introduce the representation of the multimedia data, and then present the calculation for obtaining the multi-order statistical features. Finally, we will provide the kernelized operation for the multi-order statistical features.

A. Sample Representation

We are given a training set from two different modalities: image modality $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ with class labels $\{l_1^x, l_2^x, \cdots, l_n^x\}$, text modality $\mathcal{Y} = \{y_1, y_2, \cdots, y_n\}$ with class labels $\{l_1^y, l_2^y, \cdots, l_n^y\}$, where n denotes the number of samples. Each pair $\{x_i, y_i\}$ belongs to the same class and expresses the same semantic content.

1) Image Representation: In many previous works, the two-dimensional image matrices are transformed into one-dimensional feature vectors by adopting the BoVW model. The resulting image vectors effectively reflect the concentration of visual words, while the structure information of the feature spaces will be ignored. In fact, the structure information is useful to characterize discriminative content of images such that images from different classes can be effectively discriminated.

Many methods have been proposed to exploit the structure information of image feature space. For example, Li *et al.* [33] propose the region covariance to characterize the distribution of local regions within an image, by which they achieve superior performance on object tracking.

Different from region covariance in which each pixel inside the image region serves as a sample, MDSSL first divides all images into many local patches with constant size and then

the local feature descriptors are computed by conducting dense SIFT algorithm [23], [34]. By this way, every image is modeled as a set of local feature descriptors. Specifically, the i-th image x_i is represented as $S_{ij}^x, j = 1, \ldots, |k_i^x|$, where $|k_i^x|$ denotes cardinality of S_i^x. To characterize the structure information of feature space, each image is modeled as a 2nd-order covariance matrix rather than simply calculating their term frequency of visual words.

2) Text Representation: Most existing cross-modal methods represent text documents by adopting BoW model. Since the models is realized based on calculating the term frequency, the semantic relation among words are usually ignored such that many relevant words are taken as the individual terms. For example, both "football" and "soccer" are used to describe the same semantic content though their spellings are completely different. The semantic relation between "football" and "soccer" may be ignored by using the model. Besides, BoW representation ignores the word orders among words of a sentence or an article. It is a common sense that a text document usually consists of a set of ordered words, and the order information among these words is also useful to characterize a text document.

We adopt the *word2vec* [24] model to effectively exploit the semantic relation and the order information among words. The *Word2vec* model can efficiently learn the high quality vector representation of words from a large amounts of unstructured text data. For example, the vector of "football" is closer to "soccer" than to any other word vectors in the feature space. In this paper, since each text document is composed of the ordered words, it is represented as a matrix by the ordered word vectors of these words. specifically, the i-th text y_i is represented as $S_{ij}^y, j = 1, \ldots, |k_i^y|$, where $|k_i^y|$ denotes the word number of S_i^y.

B. Multiorder Statistical Features

After sample representations, for every sample (i.e., image or text), we calculate the 0th-order, 1st-order and 2nd-order statistics as its features. The multi-order statistics have been successfully utilized to characterize the structure of image set [35]. We believe that the multi-order statistical features can also faithfully characterize the feature spaces of different modalities from the different perspectives. In other words, the multi-order statistical features provide the complementary information for each multimedia data such that they can be used as features of each multimedia data. Therefore, we compute the multi-order statistics and take them as the sample' features. Specially, the multi-order statistics of two modalities are calculated as follows:

1) Zeroth-Order Statistics: For images, we use the local feature descriptors of all images to learn a codebook by adopting k-means algorithm. Then the i-th image is quantized into a high-dimensional histogram feature vector h_i^x by using the BoVW model. As for texts, we also adopt k-means algorithm to learn a codebook based on the word vectors learned from *word2vec* model. Similarly, the i-th text is quantized into a high-dimensional histogram feature vector h_i^y. The histogram vectors reflect the distribution of the key words from the codebook and can be seen as the 0th-order statistical features.

2) First-Order Statistics: The mean vector m_i^x of the i-th image is computed as

$$m_i^x = \frac{1}{k_i^x} \sum_{j=1}^{k_i^x} S_{ij}^x. \tag{1}$$

Similarly, we also obtain the mean vector m_i^y for the i-th text. The mean vector roughly reflects the averaged position of the feature vector in the high-dimensional space.

3) Second-Order Statistics: The covariance matrix C_i^x of the i-th image is computed as

$$C_i^x = \frac{1}{|k_i^x| - 1} \sum_{j=1}^{k_i^x} (S_{ij}^x - m_i^x)(S_{ij}^x - m_i^x)^T. \tag{2}$$

Likewise, we also compute the covariance matrix C_i^y for the i-th text. The diagonal entries of covariance matrix reflect the variance of each individual item of feature vector, and the non-diagonal entries of covariance matrix reflect the correlations of the different items of feature vector. There are two reasons for selecting the covariance matrix as a sample's feature [36]. First, since covariance matrix doesn't assume the distribution of local feature descriptors, the sample with any number of local features can obtain a natural representation. Generally speaking, if two samples belong to the same category, their local feature descriptors encode the same semantic information. Then the local feature descriptors of two samples are close to each other in the high-dimensional feature space, so the corresponding entries of their covariance matrices are also close to each other. Therefore, the covariance based representation can effectively discriminate the samples from the different classes by encoding the feature correlation information specific to each class. Second, as the statistics of all the local features within a sample, the covariance matrix can largely filter out the noise-corrupting local feature descriptors by an average filter during the covariance computation.

After computing the 0th-order, 1st-order and 2nd-order statistics, the i-th image is denoted as (h_i^x, m_i^x, C_i^x) and the i-th text denoted as (h_i^y, m_i^y, C_i^y). It is easy to know that h_i^x, m_i^x, h_i^y and m_i^y are the points in the *Euclidean* spaces $\mathbb{R}^{h_x}$, $\mathbb{R}^{m_x}$, $\mathbb{R}^{h_y}$ and $\mathbb{R}^{m_y}$ respectively, while C_i^x and C_i^y in the *Riemannian* manifolds $\mathcal{M}_x$ and $\mathcal{M}_y$, respectively.

C. Kernelized Operation for Multiorder Statistics

The covariance matrices from different modalities lie on different *Riemannian* manifolds, so it is difficult to measure their similarity directly. Aiming at the problem, we embed the *Riemannian* manifolds $\mathcal{M}_x$ and $\mathcal{M}_y$ into the high dimensional kernel spaces. This embedding not only accounts for the geometry of the *Riemannian* manifold, but also adheres to the *Euclidean* geometry [37], [38]. Specially, the space embedding for image modality $\varphi^{(2)}$ and text modality $\phi^{(2)}$ are respectively calculated as follows:

$$\begin{aligned} \varphi_{i,j}^{(2)} &= \exp\big(-d_{C_i^x, C_j^x}^2 / 2\sigma_{x^{(2)}}^2\big) \\ \phi_{i,j}^{(2)} &= \exp\big(-d_{C_i^y, C_j^y}^2 / 2\sigma_{y^{(2)}}^2\big) \end{aligned} \tag{3}$$

where $\sigma_{x^{(2)}}$ and $\sigma_{y^{(2)}}$ are kernel widths specified from the mean of distances d_{C^x,C^x} and d_{C^y,C^y}, $d_{C_i^z,C_j^z}$ (z indicates x or y) measures the Log-Euclidean Distance (LED) between the covariance matrices [38]

$$d_{C_i^z,C_j^z} = \|\log(C_i^z) - \log(C_j^z)\|_F. \tag{4}$$

To keep consistency with the 2nd-order statistical features, we also calculate the kernelized features for the other two order statistical features. After the kernelized operation, each sample is represented as 0th-order, 1st-order, 2nd-order kernelized features. Let $X_r = [\varphi_1^{(r)}, \varphi_2^{(r)}, \cdots, \varphi_m^{(r)}]$ be the r th-order kernelized feature set of all training images, and the r th-order kernelized feature set of all training texts is denoted as $Y_r = [\phi_1^{(r)}, \phi_2^{(r)}, \cdots, \phi_m^{(r)}]$. $\varphi_i^{(r)}$ and $\phi_i^{(r)}$ are the rth corresponding kernelized features of $x_i^{(r)}$ and $y_i^{(r)}$. Concretely, the multi-order kernelized features of the i-th image is represented as $\{\varphi_i^{(0)}, \varphi_i^{(1)}, \varphi_i^{(2)}\}$, and $\{\phi_i^{(0)}, \phi_i^{(1)}, \phi_i^{(2)}\}$ represents the kernelized features of the i-th text.

IV. Discriminative Structured Subspace Learning

In this section, we first present a novel framework to learn the discriminative structured subspace, which can effectively integrate multiple distance metrics. Then, an iterative method is designed to optimize the proposed framework.

A. The Common Subspace

As is well known, the kernelized features lie on different kernel spaces. It is still difficult to measure the similarity between the heterogeneous data. Therefore, we learn the multiple transformations $\mathbf{U} = [U_0, U_1, U_2]$ and $\mathbf{V} = [V_0, V_1, V_2]$, which map the multi-order kernelized spaces into a common subspace. Then the distance between image φ_i and text ϕ_j can be calculated as

$$d(\varphi_i, \phi_j) = \sum_{r=0}^{2} \alpha_r \|U_r^T \varphi_i^{(r)} - V_r^T \phi_j^{(r)}\|_F \tag{5}$$

where α_0, α_1 and α_2 are the balancing parameters and their sum is set to 1.

In the learned common subspace, the distance between the heterogeneous intra-class samples should be minimized, and the inter-class samples should be maximized simultaneously.[1] Likewise, the homogeneous intra-class and inter-class information should also be ensured, which will segregate the different classes into the different regions in the subspace. Thus, the underlying intrinsic geometric structure will be consistent with that in the original space.

B. Objective Function

The objective function $\mathcal{O}(\mathbf{U}, \mathbf{V})$ of MDSSL is defined to learn the optimal transformations $\mathbf{U}$ and $\mathbf{V}$

$$\min_{\mathbf{U},\mathbf{V}} \{D(\mathbf{U}, \mathbf{V}) + \lambda_1 G(\mathbf{U}, \mathbf{V}) + \lambda_2 T(\mathbf{U}, \mathbf{V})\} \tag{6}$$

[1]In this paper, the intraclass samples means that samples belong to the same semantic class. The interclass samples denotes that samples are assigned with the different class labels.

where $D(\mathbf{U}, \mathbf{V})$ is the distance constraint defined on the sets of similarity and dissimilarity constraints; $G(\mathbf{U}, \mathbf{V})$ is the geometry structure constraint; and $T(\mathbf{U}, \mathbf{V})$ is the regularizer defined on the target transformations $\mathbf{U}$ and $\mathbf{V}$. $\lambda_1 > 0$ and $\lambda_2 > 0$ are the tradeoff parameters.

1) Distance Constraint: The classic cross-modal methods only focus on learning the transformations from a single order statistical features. For integrating the multi-order statistical features, they can separately learn the local metrics for each order statistical features and then combine the multiple local metrics by manually setting weights for each local metric. However, it is difficult to specify the proportion of each metric by the manual setting. To exploit the interaction of the different metrics and take advantage of the information of the different statistics, in this paper, we effectively integrate the multiple local metrics to exploit more discriminative information for matching the heterogeneous data.

$$D(\mathbf{U}, \mathbf{V}) = \sum_{r=0}^{2} \alpha_r D(U_r, V_r) \tag{7}$$

Most existing related methods only take account of the positive pairs as the constraints [10], which effectively enhance the similarities among the samples belonging to the same class. From the viewpoint of retrieval, it is equally important to minimize the variance of the intra-class samples and maximize the separability of the inter-class samples. Thus, we take both the positive and negative pairs as the constraints to minimize the distances of samples within the same class and maximize the distances of samples belonging to the different classes simultaneously. In MDSSL, we adopt the classical sum of the squared distances to define each local metric

$$D(U_r, V_r) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Z(i,j) \left\| U_r^T \varphi_i^{(r)} - V_r^T \phi_j^{(r)} \right\|^2 \tag{8}$$

$$Z(i,j) = \begin{cases} 1 & \text{if } l_i^x = l_j^y \\ -1 & \text{if } l_i^x \neq l_j^y \end{cases} \tag{9}$$

where $Z(i,j)$ indicates whether the heterogeneous points $\varphi_i^{(r)}$ and $\phi_j^{(r)}$ are relevant or irrelevant inferred from their class labels. To balance the effect of the similarity and dissimilarity constraints, we normalize Z by the number of pairs with the same (different) class labels.

2) Geometry Constraint: To ensure that the neighboring data in the original feature space are still close to each other in the common subspace, we introduce the constraint $G(\mathbf{U}, \mathbf{V})$ to preserve the data distribution

$$G(\mathbf{U}, \mathbf{V}) = \sum_{r=0}^{2} G(U_r, V_r) = \sum_{r=0}^{2} (G_x(U_r) + G_y(V_r)) \tag{10}$$

where $G_x(U_r)$ and $G_y(V_r)$ are used to preserve the data distributions of images' and texts' r-th order kernelized feature, respectively.

The graph construction method has been widely used to preserve the local structure by constructing the relations for the positive pairs [39]. In this paper, we construct the relations for

both the positive and negative pairs. Specially, similar to Linear Discriminant Analysis (LDA), we design $G_x(U_r)$ as

$$G_x(U_r) = G_x^w(U_r) - G_x^b(U_r) \tag{11}$$

where $G_x^w(U_r)$ is used to construct the relations for the nearest intra-class samples. It ensures that the close samples within the same class are more close after the transformation. Conversely, $G_x^b(U_r)$ ensures that the close samples belonging to the different classes are separated as far as possible in the common subspace.

In practice, given a query, many retrieved samples usually have the same semantic label or different semantic labels with it. Hence, each query usually has a lot of inter-class samples and intra-class samples. Considering the effectiveness and efficiency, we apply the conventional k-nearest neighbor method to compute $G_x^w(U_r)$ and $G_x^b(U_r)$. By this way, we can construct the relations for the nearest intra-neighbors and nearest inter-neighbors. $G_x^w(U_r)$ can be simplified to the following formulation after some algebraic manipulations:

$$G_x^w(U_r) = \frac{1}{2}\sum_{i=1}^{n}\sum_{p=1}^{k_1}\left\|U_r^T\varphi_i^{(r)} - U_r^T\varphi_{ip}^{(r)}\right\|^2 A_{ip} \tag{12}$$

$$A_{ip} = \begin{cases} \exp\left(-\|\varphi_i^{(r)} - \varphi_{ip}^{(r)}\|^2/\sigma_{wr}^2\right) & \text{if } \varphi_{ip}^{(r)} \in N_{intra}^{k_1}\left(\varphi_i^{(r)}\right) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $N_{intra}^{k_1}(\varphi_i^{(r)})$ denotes the k_1-nearest intra-neighbors of $\varphi_i^{(r)}$, $\varphi_{ip}^{(r)}$ represents the p-th nearest intra-neighbors of $\varphi_i^{(r)}$ and A is the affinity matrix to characterize the similarity between the nearest intra-class samples. σ_{wr} is the kernel widths obtained by calculating the mean of distances of all the nearest intra-neighbors.

Similarly, we can simplify G_x^b as follows:

$$G_x^b(U_r) = \frac{1}{2}\sum_{i=1}^{n}\sum_{q=1}^{k_2}\left\|U_r^T\varphi_i^{(r)} - U_r^T\varphi_{iq}^{(r)}\right\|^2 B_{iq} \tag{14}$$

$$B_{iq} = \begin{cases} \exp\left(-\|\varphi_i^{(r)} - \varphi_{iq}^{(r)}\|^2/\sigma_{br}^2\right) & \text{if } \varphi_{iq} \in N_{inter}^{k_2}\left(\varphi_i^{(r)}\right) \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $N_{inter}^{k_2}(\varphi_i^{(r)})$ denotes the k_2-nearest inter-neighbors of $\varphi_i^{(r)}$. $\varphi_{iq}^{(r)}$ denotes the qth nearest inter-neighbors of $\varphi_i^{(r)}$. B is the affinity matrix to characterize the similarity between the nearest inter-class samples. σ_{wr} is the kernel widths obtained by calculating the mean of distances of all the nearest inter-neighbors.

Similar to $G_x^w(U_r)$ and $G_x^b(U_r)$, we can obtain $G_y^w(V_r)$ and $G_y^b(V_r)$ by replacing U_r with V_r, respectively.

3) Transformation Regularization: $T(\mathbf{U},\mathbf{V})$ is defined to control the scale of transformations and reduce overfitting. Its formulation is defined as follows:

$$T(\mathbf{U},\mathbf{V}) = \frac{1}{2}\sum_{r=0}^{2}\left(\left\|U_r^T X_r\right\|_F^2 + \left\|V_r^T Y_r\right\|_F^2\right). \tag{16}$$

C. Iterative Optimization

To obtain the optimal solution of (6), we propose an unified framework to minimize the objective function by iterative strategy. We first initialize $\mathbf{U}$, $\mathbf{V}$ and $\alpha = [\alpha_0, \alpha_1, \alpha_2]$. Then, we alternately update $\mathbf{U}$ and $\mathbf{V}$ in each iteration. After obtaining $\mathbf{U}$ and $\mathbf{V}$, we adopt the *Lagrangian* function to optimize α.

1) Initialization: In this paper, we adopt the within-class and between-class analysis to initialize $\mathbf{U}$ and $\mathbf{V}$. This strategy ensures the discriminative structures of the transformations at first and reduces the number of iterations.

To conduct this analysis, we compute the within-class and between-class distance constraints $D^w(U_r, V_r)$ and $D^b(U_r, V_r)$ by replacing Z with Z^w and Z^b in (9). Z^w and Z^b are computed as

$$Z^w(i,j) = \begin{cases} 1 & \text{if } l_i^x = l_j^y \\ -1 & \text{if } l_i^x \neq l_j^y \end{cases} \tag{17}$$

$$Z^b(i,j) = \begin{cases} -1 & \text{if } l_i^x = l_j^y \\ 1 & \text{if } l_i^x \neq l_j^y. \end{cases} \tag{18}$$

For the geometry constraint, we directly use the within-class and between-class templates $G_x^w(U_r)$, $G_x^b(U_r)$, $G_y^w(V_r)$ and $G_y^b(V_r)$ in (10). Then, U_r and V_r can be initialized by minimizing the within-class templates while maximizing the between-class templates

$$\begin{aligned} &\min_{U_r,V_r}\left\{D^w(U_r,V_r) + \lambda_1(G_x^w(U_r) + G_y^w(V_r))\right\} \\ &\text{s.t. } D^b(U_r,V_r) + \lambda_1(G_x^b(U_r) + G_y^b(V_r)) = 1. \end{aligned} \tag{19}$$

The objective function is a standard generalized eigenvalue problem that can be solved by adopting any eigensolver.

2) Update $\mathbf{U}$*:* Differentiating $\mathcal{O}(\mathbf{U},\mathbf{V})$ with respect to U_r, we have the following equation:

$$\begin{aligned} \frac{\partial\mathcal{O}(\mathbf{U},\mathbf{V})}{\partial U_r} &= \alpha_r X_r Q_x X_r^T U_r - \alpha_r X_r Z Y_r^T V_r \\ &\quad + \lambda_1(J_{xr}^w - J_{xr}^b)U_r + \lambda_2 X_r X_r^T U_r \end{aligned} \tag{20}$$

where Q_x is a diagonal matrix with $Q_x(i,i) = \sum_{j=1}^{n} Z(i,j)$, J_{xr}^w and J_{xr}^b are two intermediate variables, which are used to define the intra-class and inter-class relations, respectively

$$\begin{aligned} J_{xr}^w &\triangleq \sum_{i=1}^{n}\sum_{p=1}^{k_1}\left(\varphi_i^{(r)} - \varphi_{ip}^{(r)}\right)\left(\varphi_i^{(r)} - \varphi_{ip}^{(r)}\right)^T A_{ip} \\ J_{xr}^b &\triangleq \sum_{i=1}^{n}\sum_{q=1}^{k_2}\left(\varphi_i^{(r)} - \varphi_{iq}^{(r)}\right)\left(\varphi_i^{(r)} - \varphi_{iq}^{(r)}\right)^T B_{iq}. \end{aligned} \tag{21}$$

Then, setting (20) to 0, we can obtain

$$U_r = \left(X_r Q_x X_r^T + \frac{\lambda_1}{\alpha_r}(J_{xr}^w - J_{xr}^b) + \frac{\lambda_2}{\alpha_r}X_r X_r^T\right)^{-1} X_r Z Y_r^T V_r. \tag{22}$$

3) Update $\mathbf{V}$*:* Similarly, differentiating $\mathcal{O}(\mathbf{U}, \mathbf{V})$ with respect to V_r and setting it to zero, we obtain

$$V_r = \left(Y_r Q_y Y_r^T + \frac{\lambda_1}{\alpha_r}(J_{yr}^w - J_{yr}^b) + \frac{\lambda_2}{\alpha_r} Y_r Y_r^T \right)^{-1} Y_r Z X_r^T U_r \tag{23}$$

where Q_y is a diagonal matrix with $Q_y(j,j) = \sum_{i=1}^{n} Z(i,j)$, J_{yr}^w and J_{yr}^b are calculated as (21).

4) Update α*:* To automatically exploit the complementary information of different statistical features, we modify α_r as α_r^β, where $\beta > 1$. Then we reconstruct the objective function by adopting the *Lagrangian* function such that α_r can be solved automatically in each iteration

$$\begin{aligned} \hat{\mathcal{O}}(\alpha, \eta) = & \sum_{r=0}^{2} \alpha_r^\beta D(U_r, V_r) + \eta \left(\sum_{r=0}^{2} \alpha_r - 1 \right) \\ & + \lambda_1 G(\mathbf{U}, \mathbf{V}) + \lambda_2 T(\mathbf{U}, \mathbf{V}) \end{aligned} \tag{24}$$

where η is the *Lagrangian* multiplier. In order to get the optimal α_r, we set the derivative of $\hat{\mathcal{O}}(\alpha, \eta)$ with respect to α_r and η to zero. Then we obtain α_r as follows:

$$\alpha_r = \frac{(1/D(U_r, V_r))^{1/(\beta-1)}}{\sum_{r=0}^{2}(1/D(U_r, V_r))^{1/(\beta-1)}}. \tag{25}$$

To search an optimal solution, we alternate the above updates of $\mathbf{U}$, $\mathbf{V}$ and α. The convergence criterion used in our experiments is that the performance on the validation set decreases or $|\mathcal{O}_{t-1} - \mathcal{O}_t| \leq 0.001$, where $\mathcal{O}_t$ is the value of the objective function in the t-th iteration.

D. Computational Complexity

Assuming d_x is the dimension of image's local descriptor, d_y is the dimension of word vector. The asymptotic time complexity of MDSSL is $O(n \times \mathrm{d}_x^3 + n \times \mathrm{d}_y^3 + \mathcal{N} \times n^3)$. $O(n \times \mathrm{d}_x^3)$ and $O(n \times \mathrm{d}_y^3)$ are the cost for computing the kernelized features of images and texts, respectively. In each iteration, the computational complexity, according to (22) and (23), is $O(n^3)$ for matrix multiplication, inverse and eigenvalue decomposition since the dimension of the kernelized feature is n. After $\mathcal{N}$ iterations, the total computational complexity is $O(\mathcal{N} \times n^3)$. It is also noted that MDSSL suffers high cost of computing kernelized features, while MDSSL would be convergent after several iterations in the stage of optimization.

V. Experimental Results

In this section, we present extensive experiments to demonstrate the effectiveness of the proposed method for text-image retrieval, i.e., image-query-texts, text-query-images and their average. We evaluate and compare different methods on two publicly available datasets: Wiki [5] and NUS-WIDE [40].

A. Datasets

Wiki[2] contains 2,866 articles generating from Wikipedia's featured articles [5]. Each article consists of a pair of image and text description, and it is categorized into 10 semantic classes. We randomly choose 1,500 pairs of the data for training, 500 pairs for validation and 866 pairs for testing.

The NUS-WIDE dataset[3] consists of 269,648 paired samples with 81 concepts [40]. Each image with its annotated tags can be treated as an image-text pair. We randomly select 6,664 images that have at least one tag and one concept from the 10 largest concepts, which are regarded as the categories in this paper. Then 2,664 paired samples are used for training, 2,000 for validation and 2,000 for testing.

For the two datasets, we take the same strategy to extract image features. The dense SIFT features are extracted for each image at first. For 0th-order statistics, each image from the Wiki (NUS-WIDE) dataset is quantized into a 1000 (500) dimensional histogram feature vector by adopting BoVW. Since the dimension of the SIFT features is 128, we also obtain a 128-dimensional mean vector and a 128×128 covariance matrix for each image.

As for the textual features, we first use the Google corpus to train the word vectors by adopting the *word2vec* model [24]. Then, each textual document is represented as a set of vectors on the Wiki dataset. The tags associated with each image are represented as a set of vectors on the NUS-WIDE dataset. Finally, the multi-order statistics are computed as the feature representation for texts. Specially, we extract the 0th-order features similar to the BoVW model rather than the BoW model. The dimensions of the multi-order statistical features are determined by empirical analysis. We obtain the histogram feature vectors with 1500-dimension and 1000-dimension on the Wiki and NUS-WIDE dataset, respectively. Specifically, we train the 100-dimensional word vectors on the Google corpus. Then each text also obtains a 100-dimensional mean vector and a 100×100 covariance matrix.

B. Experimental Settings

MDSSL is compared with CCA [1], SCM [5], ml-CCA [8], LGCFL [19], LCFS [20], Bi-CMSRM [15], GMLDA, GMMFA [10], 3-view CCA [11], cluster-CCA [12] and BITR [18].

CCA learns a common subspace where the correlation between two modalities are maximized. Comparing with CCA can validate MDSSL's ablity on learning the useful latent space. SCM, ml-CCA, GMLDA, GMMFA, 3-view CCA and cluster-CCA are CCA-based methods. SCM adopts logistic regression in the CCA projected coefficient space and obtains the posterior probabilities assigned to all semantic classes. SCM and MDSSL use kernel function to obtain new feature representation, so comparing with SCM can validate the effectiveness of the proposed metric framework. As the supervised extensions of CCA, the other CCA-based methods exploit the label information for learning discriminant latent space. We select these methods to validate the ablity on utilizing the label information. LGCFL and LCFS use the label space as a linkage to learn a coupled of mappings by optimizing the labeling ap-

[2][Online]. Available: http://www.svcl.ucsd.edu/projects/crossmodal/

[3][Online]. Available: http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

TABLE I
PERFORMANCE COMPARISON IN TERMS OF MAP@R SCORES ON THE WIKI DATASET WITH DIMENSIONALITY OF LATENT SPACE S EQUALS TO 8

Methods \ Tasks	R = 5			R = 50			R = all		
	Text query	Image query	Average	Text query	Image query	Average	Text query	Image query	Average
CCA	0.3938	0.3110	0.3524	0.2921	0.2643	0.2782	0.1723	0.1927	0.1825
SCM	0.4325	0.2977	0.3651	0.3289	0.2494	0.2892	0.2017	0.2099	0.2058
LCFS	0.5126	0.2755	0.3941	0.3664	0.2610	0.3137	0.2087	0.2517	0.2302
BITR	0.4017	0.2916	0.3404	0.2879	0.2583	0.2731	0.1792	0.2272	0.2032
LGCFL	0.4152	0.3012	0.3582	0.3321	0.2879	0.3100	0.2099	0.2768	0.2434
ml-CCA	0.5625	0.2896	0.4204	0.3838	0.2603	0.3221	0.2073	0.2634	0.2354
GMLDA	0.4871	0.2691	0.3781	0.3463	0.2379	0.2921	0.1939	0.2483	0.2211
GMMFA	0.5017	0.2694	0.3856	0.3477	0.2391	0.2934	0.2034	0.2469	0.2252
Bi-CMSRM	0.5643	0.2799	0.4221	0.3715	0.2695	0.3205	0.2093	0.2589	0.2341
cluster-CCA	0.5007	0.2924	0.3965	0.3353	0.2613	0.2983	0.1857	0.2405	0.2131
3-view CCA	0.5263	0.2956	0.4110	0.3474	0.2674	0.3074	0.1942	0.2509	0.2225
MDSSL$^{(0)}$	0.5844	0.3182	0.4513	0.4341	0.3016	0.3679	0.2306	0.3062	0.2684
MDSSL$^{(1)}$	0.4661	0.2567	0.3614	0.3190	0.2528	0.2859	0.1897	0.2759	0.2328
MDSSL$^{(2)}$	0.6003	0.3484	0.4744	0.4437	0.3311	0.3874	0.2556	0.3281	0.2919
MDSSL	**0.6702**	**0.3677**	**0.5190**	**0.4888**	**0.3473**	**0.4181**	**0.2851**	**0.3517**	**0.3184**

Both directions of retrieval tasks are reported. The results shown in boldface are the best results.

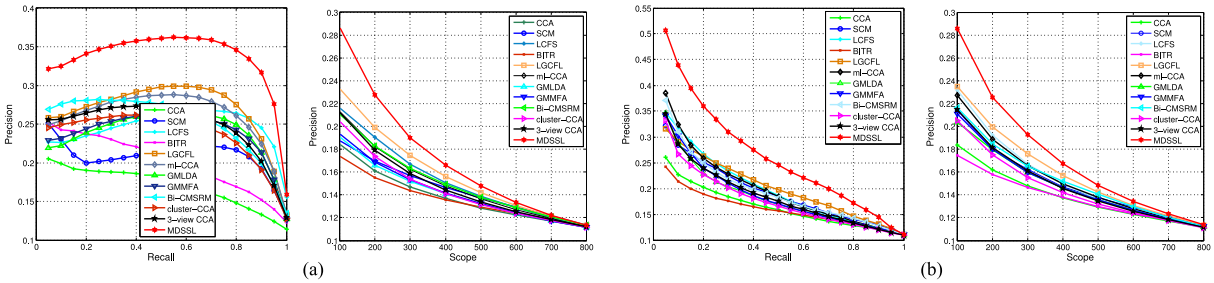


Fig. 2. Precision-recall curves and precision-scope curves for text-image retrieval on the Wiki dataset. (a) Image-query-texts. (b) Text-query-images.

TABLE II
PERFORMANCE COMPARISON IN TERMS OF MAP@R SCORES ON THE NUS-WIDE DATASET WITH DIMENSIONALITY OF LATENT SPACE S EQUALS TO 8

Methods \ Tasks	R = 5			R = 50			R = all		
	Text query	Image query	Average	Text query	Image query	Average	Text query	Image query	Average
CCA	0.4690	0.3200	0.3945	0.3331	0.2749	0.3040	0.1824	0.2178	0.2001
SCM	0.5176	0.3137	0.4157	0.3736	0.2765	0.3251	0.2149	0.2378	0.2264
LCFS	0.5417	0.2831	0.4124	0.4418	0.2931	0.3675	0.2346	0.3567	0.2957
BITR	0.4390	0.3636	0.4013	0.3373	0.3243	0.3308	0.2143	0.2543	0.2343
LGCFL	0.6186	0.3171	0.4676	0.4515	0.3002	0.3759	0.2503	0.3732	0.3118
ml-CCA	0.5404	0.3013	0.4208	0.4238	0.2926	0.3582	0.2367	0.3617	0.2992
GMLDA	0.5367	0.2875	0.4121	0.4322	0.2850	0.3586	0.2450	0.3638	0.3044
GMMFA	0.5279	0.2949	0.4114	0.4317	0.2832	0.3675	0.2501	0.3680	0.3091
Bi-CMSRM	0.4373	0.2662	0.3518	0.3313	0.2515	0.2914	0.1959	0.2620	0.2290
cluster-CCA	0.5407	0.2817	0.4112	0.3874	0.2687	0.3280	0.2091	0.3245	0.2669
3-view CCA	0.5745	0.2963	0.4354	0.4259	0.2842	0.3550	0.2290	0.3504	0.2897
MDSSL$^{(0)}$	0.5685	0.3280	0.4483	0.4811	0.3095	0.3953	0.2704	0.3953	0.3329
MDSSL$^{(1)}$	0.5358	0.2644	0.4001	0.3588	0.2625	0.3107	0.2124	0.3527	0.2826
MDSSL$^{(2)}$	0.7964	0.4358	0.6161	0.6660	0.4313	0.5586	0.4047	0.5138	0.4593
MDSSL	**0.8160**	**0.4389**	**0.6275**	**0.6673**	**0.4354**	**0.5514**	**0.4079**	**0.5218**	**0.4649**

Both directions of retrieval tasks are reported. The results shown in boldface are the best results.

proximation error between the given data and labels. By this way, they can learn the more discriminative low-dimensional feature representation. The two methods are selected to validate the discrimination of the low-dimensional feature representation. Bi-CMSRM and BITR are the learning to rank methods, which aim at optimizing the top of ranking. The top-ranked performance of MDSSL can be validated by comparing with it.

Additionally, since MDSSL uses the multi-order statistical features to represent each sample, we also report the performance of MDSSL$^{(0)}$, MDSSL$^{(1)}$ and MDSSL$^{(2)}$. They only use one order statistical features for cross-modal retrieval. For example, MDSSL$^{(0)}$ denotes that only the 0th-order statistics are used to match images and texts.

For the evaluation, we use mean average precision (MAP) [5] as the performance measures. MAP@R measures MAP scores at

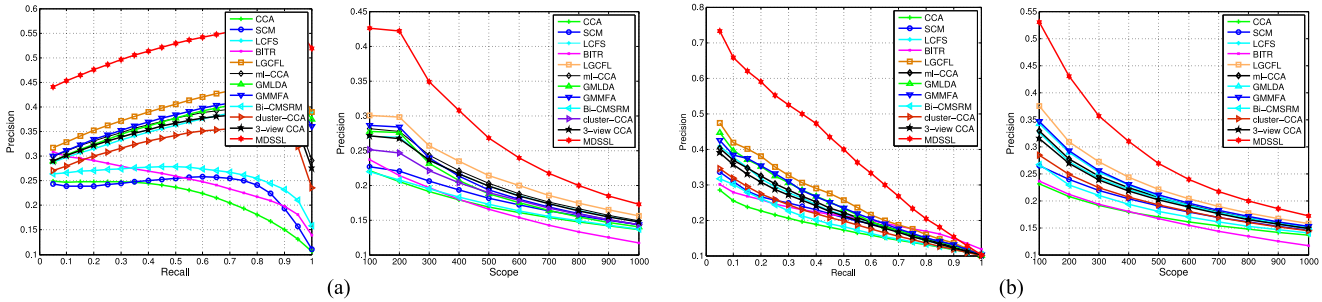


Fig. 3. Precision-recall curves and precision-scope curves for text-image retrieval on the NUS dataset. (a) Image-query-texts. (b) Text-query-images.

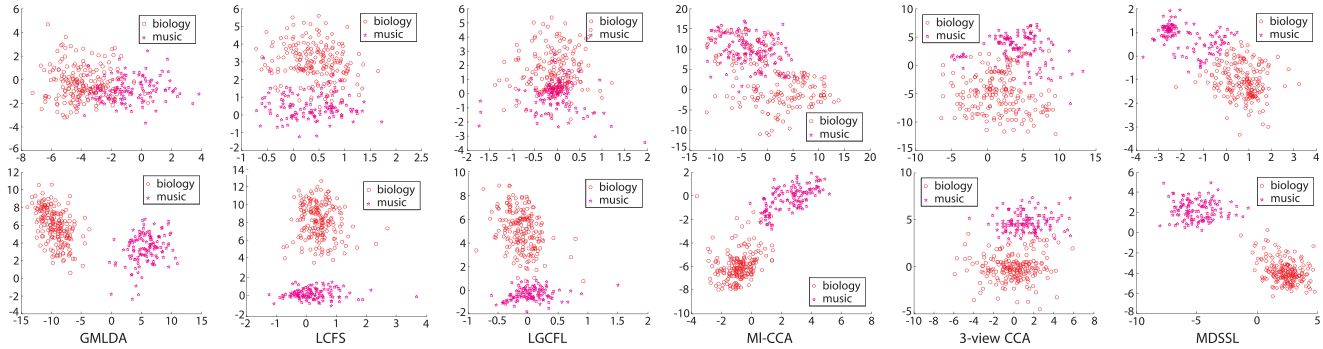


Fig. 4. Low-dimensional mapping of images and texts from "biology" and "music" classes on Wiki dataset. The top row shows the mapping for image modality, and the bottom shows the mapping for text modality.

fixed number of retrieved samples, and we set R to 50 for the top 50 retrieved samples and R to all for all the retrieved samples. Besides, to give a pictorial demonstration of the performance, we also display the precision-recall curve [5] and precision-scope curve [41] for all the methods. The scope is specified by the number of the top-ranked samples when the retrieved samples are ranked according to the similarities between them and the query.

For all methods, we use the optimal parameters settings tuned by a parameter validation process except for the specified values. For CCA, SCM, GMLDA and GMMFA, principal component analysis (PCA) is performed on the original features to remove the redundant features, where 95% feature energy is preserved. The proposed method uses the following parameter settings: $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, $\beta = 2$, $k_1 = 40$, $k_2 = 300$. On both datasets, we repeat the experiments 10 times by randomly selecting training/valiadtion/testing combinations, and show the average MAP of the different methods. Besides, for fairness, all the compared methods adopt the 0th-order kernelized features in our experiments.

C. Results on Wiki Dataset

Table I shows the MAP scores of different methods on Wiki dataset by varying the number of retrieved samples R. From the table, we can draw the following conclusions:

First, in two retrieval tasks, the MAP scores of MDSSL$^{(0)}$ are higher than those of other comparative methods except for MDSSL$^{(2)}$ and MDSSL. For example, MDSSL$^{(0)}$ achieves 0.2306 and 0.3062 for text query and image query, while LGCFL achieves 0.2099 and 0.2768, respectively. Since all methods use the same 0th-order kernelized features, we attribute the improvement of MDSSL$^{(0)}$ to the metric framework. In MDSSL, the metric framework takes both the positive and negative pairs to constrain the distance and space structure, while LGCFL optimizes the labeling error loss but ignores the data distribution. This experiment demonstrates that our metric framework can effectively measure the similarity between the different modalities.

TABLE III
MAP SCORES BY CALCULATING TERM FREQUENCY ON THE WIKI DATASET

Methods \ Tasks	Text query	Image query	Average
CCA	0.1626	0.1811	0.1718
SCM	0.1962	0.2025	0.1993
LCFS	0.1793	0.2264	0.2028
BITR	0.1701	0.2117	0.1909
LGCFL	0.2011	0.2421	0.2216
ml-CCA	0.1906	0.2611	0.2259
GMLDA	0.1874	0.2195	0.2025
GMMFA	0.1843	0.2231	0.2037
Bi-CMSRM	0.1879	0.2142	0.2011
cluster-CCA	0.1849	0.2401	0.2125
3-view CCA	0.1892	0.2431	0.2162
MDSSL$^{(0)}$	0.2039	0.2657	0.2348

Second, the performance of MDSSL$^{(2)}$ outperforms that of MDSSL$^{(0)}$. Specially, the MAP scores of MDSSL$^{(2)}$ are 0.2556 and 0.3281 for text query and image query, respectively. Considering that the difference between MDSSL$^{(0)}$ and MDSSL$^{(2)}$ is that images and texts are represented by the different order statistical features, the results manifest that the structure in-

TABLE IV
MAP SCORES ON USING DEEP FEATURES ON THE WIKI DATASET

Methods \ Tasks	Text query	Image query	Average
CCA	0.1779	0.1999	0.1889
SCM	0.1997	0.2348	0.2173
LCFS	0.2098	0.2767	0.2433
BITR	0.1932	0.2316	0.2124
LGCFL	0.2564	0.2967	0.2766
ml-CCA	0.2117	0.2705	0.2411
GMLDA	0.2554	0.2579	0.2567
GMMFA	0.2484	0.2519	0.2502
Bi-CMSRM	0.2123	0.2601	0.2362
cluster-CCA	0.1909	0.2555	0.2232
3-view CCA	0.2091	0.2662	0.2377
MDSSL$^{(0)}$	0.2597	0.3126	0.2862

formation encoded in two-dimensional covariance matrices is beneficial for matching the heterogeneous data.

Finally, by integrating MDSSL$^{(0)}$, MDSSL$^{(1)}$ and MDSSL$^{(2)}$, the performance of MDSSL can be further improved. For MDSSL, the MAP scores of text query and image query are 0.2851 and 0.3517. The results show that fusing the multi-order statistical features can enrich the semantic information, and integrating the multi-order metrics can further exploit the complementary information among the multi-order features such that the correlations between the different modalities are enhanced.

The precision-recall and scope-precision curves on both directional retrieval are shown in Fig. 2. The curves further validate the superiority of MDSSL for cross-modal retrieval.

D. Results on NUS-WIDE Dataset

The MAP scores of all the methods are shown in Table II, and the precision-recall and precision-scope curves are reported in Fig. 3. These results show that MDSSL still outperforms all the compared methods, and the above analysis on the Wiki dataset is reasonable.

The improvement of MDSSL on the NUS dataset is as significant as that on the Wiki dataset. For example, compared with the second best result from LGCFL, MDSSL has increased to 62.96% and 39.82% in the text query and image query, respectively, while the increases are about 35.83% and 27.06% on the Wiki dataset.

We also observe that the performance of MDSSL$^{(2)}$ greatly outperforms MDSSL$^{(0)}$ and MDSSL$^{(1)}$, which is different from the Wiki dataset. The distinct difference between Wiki and NUS is the textual modality, which are article and tags respectively. This phenomenon is possibly due to that the article contains some irrelevant words. Therefore, the improvement of MDSSL is smaller than MDSSL$^{(2)}$ since the three metrics are unbalanced.

The precision-recall and scope-precision curves on both directional retrieval are shown in Fig. 3. Similar to the Wiki dataset, MDSSL has the best overall performance. In Fig. 3, many methods like LGCFL, LCFS and MDSSL achieve lower precision at low levels of recall in the task of image query. LCFS and LGCFL learn the coupled mappings by optimizing the labeling approximation error between the given data and labels. It is well known that the class labels apply more directly to texts than images, so the image query is more likely to mismatch. The distance constraint of MDSSL effectively distinguishes the inter-class samples and intra-class samples and do not optimize the ranked list. This may lead to the fact that a lot of relevant samples are pushed in the front of the ranked list but not the top of ranked list.

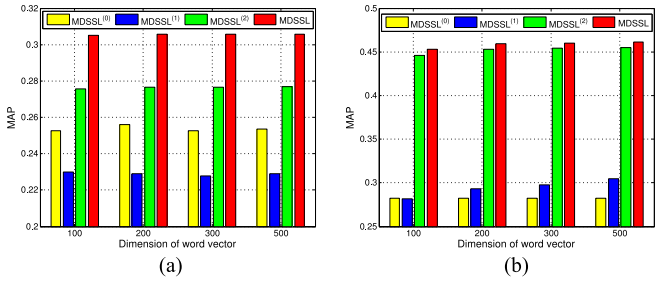


Fig. 5. Average MAP scores of MDSSL by varying dimension of word vector on both dataset. (a) Wiki. (b) NUS-WIDE.

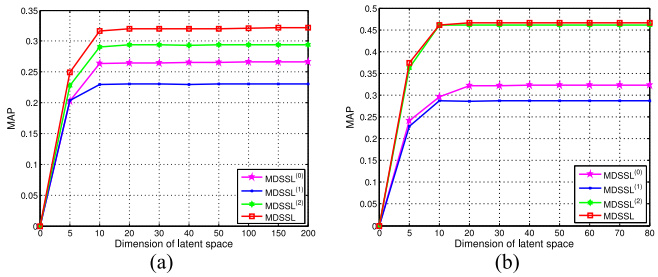


Fig. 6. MAP scores of MDSSL and its extensions on different dimensionality of latent space on both dataset. (a) Wiki. (b) NUS-WIDE.

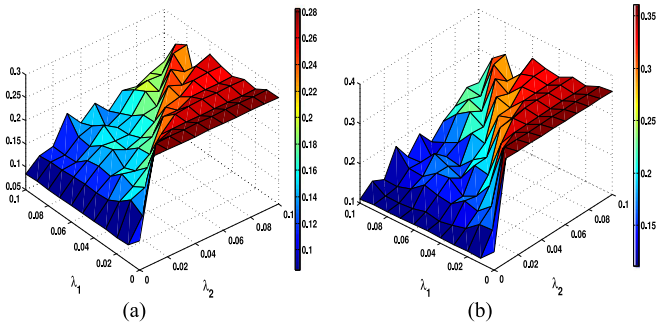


Fig. 7. MAP scores of MDSSL with different λ_1 and λ_2 on the Wiki dataset. (a) Text query. (b) Image query.

E. Results on Transformations

In this part, we show the data distribution after the low-dimensional transformation. To demonstrate this, we construct a toy dataset using the 'biology' and 'music' classes of the Wiki dataset. For both modalities, we show the 1st and 2nd most correlated components of the different methods in a two-dimensional coordinate plane. Fig. 4 shows the six best results based on the intuitive judgement. The red color represents the data distribution of the 'biology' class, and magenta color represents the 'music' class. From this figure, we can see that MDSSL

TABLE V
PERFORMANCE COMPARISON IN TERMS OF λ_1 AND λ_2 ON THE WIKI DATASET

Methods \ Tasks	λ_1=0, λ_2=0		λ_1=0.01, λ_2=0		λ_1=0, λ_2=0.1		λ_1=0.01, λ_2=0.1	
	Text query	Image query	Text query	Image query	Text query	Image query	Text query	Image query
MDSSL$^{(0)}$	0.1235	0.1371	0.1342	0.1417	0.2005	0.2781	0.2306	0.3062
MDSSL$^{(1)}$	0.1210	0.1306	0.1237	0.1354	0.1711	0.2403	0.1897	0.2759
MDSSL$^{(2)}$	0.1217	0.1354	0.1463	0.1571	0.2197	0.3005	0.2556	0.3281
MDSSL	0.1287	0.1401	0.1571	0.1804	0.2461	0.3273	0.2851	0.3517

TABLE VI
PROCESSING TIME COMPARISON (SECONDS)

Tasks \ Methods	SCM	GMLDA	GMMFA	LCFS	Bi-CMSRM	LGCFL	ml-CCA	cluster-CCA	3-view CCA	MDSSL
Training	1.35	16.38	16.39	16.71	2326.58	10.43	608.93	397.13	259.24	71.05
Test	10.83	10.61	10.48	10.44	14.26	11.01	10.71	10.95	10.69	10.46

can minimize the variance of the intra-class samples and maximize the separability of inter-class samples for two retrieval tasks, but some methods, like LCFS and LGCFL, only minimize the variance of the intra-class samples and maximize the separability of inter-class samples for text query. Both LCFS and LGCFL use the label space as a linkage, and then they ensure the low-dimensional subspaces of image and text modalities consistent with the label space. Since the class labels apply more directly to texts than images, the low-dimensional transformation of image modality is less discriminative than that of the text modality. This may also be the reason that the low-dimensional transformation of text query is superior than that of image query in all the methods. MDSSL use both the intra-class samples and inter-class samples to define the distance and geometry constraints, so samples from the different classes can be segregated into the different regions in the low-dimensional common subspace. These results validate that MDSSL is able to map the different features into a discriminative subspace such that the low-dimensional representations are enforced with the high correlation.

F. Discussion on Different Feature

In this part, we validate the effectiveness of the proposed features on matching the heterogeneous data.

According to Section III, we know that our 0th-order statistical features are different from previous BoW based representation. For fairness, we keep the same experimental setting except for calculating the 1,500-dimensional histogram features as MDSSL$^{(0)}$ in Table I. That is, all methods use the same visual features and the different textual features compared with Table I. We report the MAP scores of different methods by calculating term frequency in Table III. From this table, we conclude that the performances of all methods are lower than that in Table I, and MDSSL$^{(0)}$ still achieves the best performance. These results demonstrate that using the similarity among word vectors to calculate the histogram features is more effective than that of calculating the term frequency, and the proposed discriminative structured subspace learning is more suitable for cross-modal retrieval.

Recently, features extracted using a pre-trained deep neural network (e.g., CNN) has become quite popular for image representation in diverse vision-related tasks [32]. Following the same protocol in [42], we extract 4096-dimensional activation features from the 'fc7' layers [32] for images. For text feature, we use 100-dimensional skip-gram word vectors learned by *word2vec* and compute a mean vector of the word vectors of the words appearing in each text description. Note that the textual features are equal to the 1st-order statistical features in Table I, so this experiment uses the same textual features and different visual features with Table I. We report the results of different methods using the deep features in Table IV. From this table, we conclude that the MAP scores of all methods are improved in different degrees. For example, the average MAP scores of

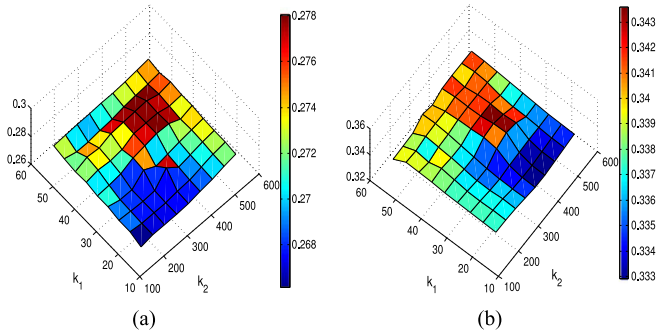


Fig. 8. MAP scores of MDSSL with different k_1 and k_2 on the Wiki dataset. (a) Text query. (b) Image query.

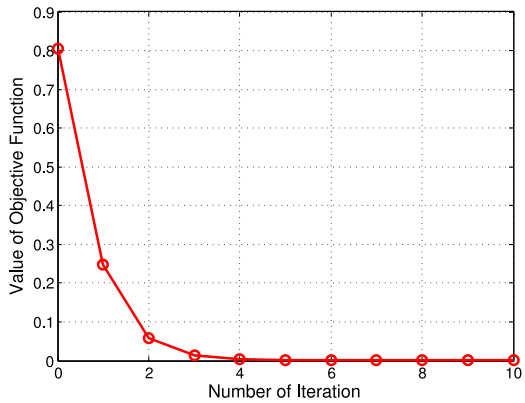


Fig. 9. Value of objective function by varying iterations on Wiki Dataset.

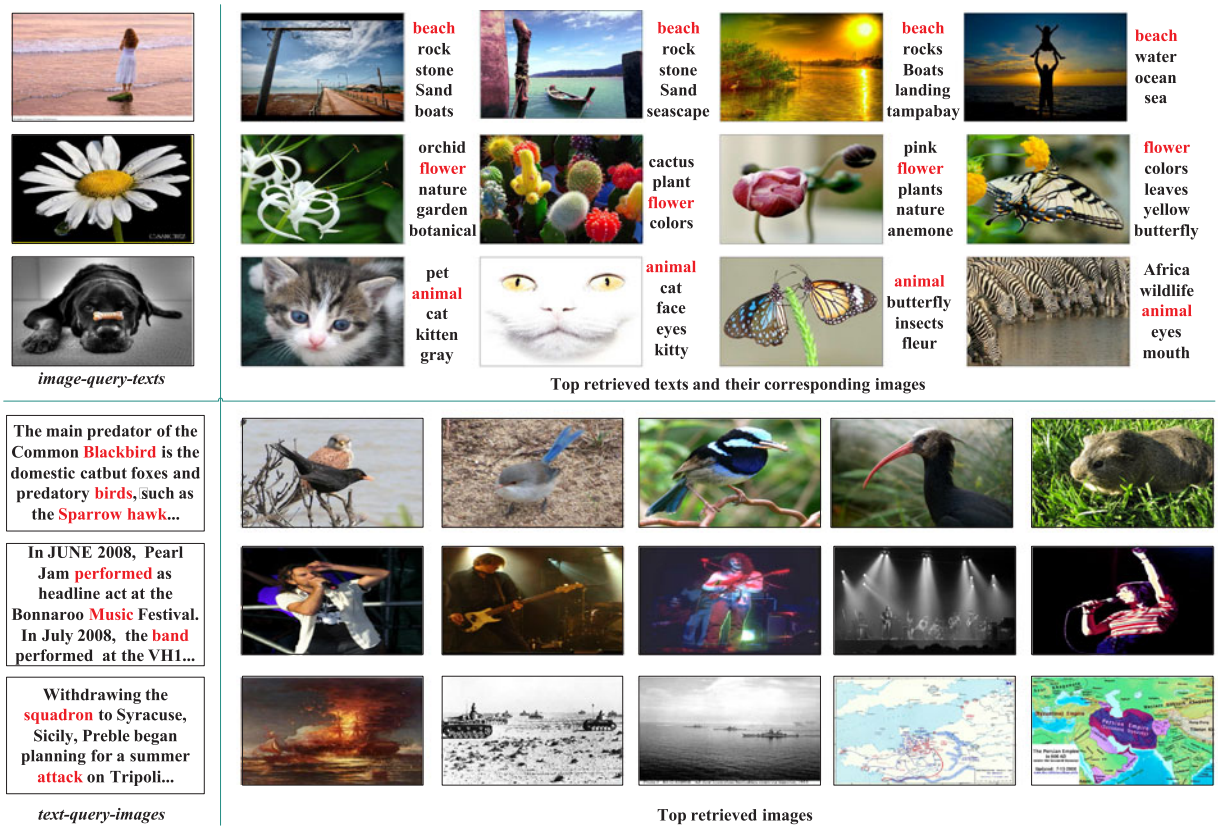


Fig. 10. Some examples of image-query-texts (in top half) on the NUS dataset and text-query-images (in bottom half) on the Wiki dataset. For each direction, we show the query and its corresponding top retrieved results by the proposed MDSSL. For the image-query-texts direction, we can find that the top retrieved texts of MDSSL are clearly relevant to the images belonging to the "beach", "flower" and "animal" category, respectively. For the text-query-images direction, given textual description about "biology", "music", "warfare", the top retrieved images of MDSSL are also relevant to the query texts.

LCFS and LGCFL are 0.2433 and 0.2766, which are about 5.7% and 13.6% higher than the results in Table I. Besides, by comparing with the results of MDSSL$^{(0)}$ in different features, we know that the covariance based representation can achieve the comparable performance with deep features. The reason is possibly due to the similarities between the CNN features and the covariance based features. For example, CNN adopts the local receptive fields to explore the semantic information of local patches in the convolutional layer, while the covariance based features use the SIFT local feature descriptors to exploit the high-order semantic information.

Finally, since the dimension of word vectors is adjustable in *word2vec* model, we do an experiment to validate its impact on the retrieval performance. From Fig. 5, we conclude that the performance is stable when the dimension of word vectors varies from 100 to 500. In this paper, we set the dimension to 100 for high efficiency.

G. Parameter Sensitivity Analysis

We conduct sensitivity analysis on parameters to test their impact on the performance of cross-modal retrieval. All sensitivity experiments are performed on the validation set of Wiki dataset and then the values are fixed throughout the experiments on both datasets.

In Fig. 6, we show the average MAP scores of MDSSL, MDSSL$^{(0)}$, MDSSL$^{(1)}$ and MDSSL$^{(2)}$ when the dimensionality of the common subspace varies from 5 to 200. From this figure, we observe that all curves have the same trend. When the dimension is increased from 0 to 10, the performance is improved in each curve. Then the performance becomes stable when the dimensions are larger than 10. These phenomenons are possibly due to the redundancy caused by high-dimensional features since the intrinsic dimension of a semantic space is usually much lower than that of original feature space.

In Fig. 7, we show the performance trend of text query and image query by varying λ_1 and λ_2. We observe that the MAP scores are in upward trend with the increase of λ_2, and downward trend with the increase of λ_1. MDSSL obtains best results when λ_1 falls into the range of $[0.005, 0.02]$ and λ_2 falls into the range of $[0.005, 0.1]$. Besides, results listed in Table V also indicate that integrating the distance, geometry and transformation constraints performs much better than that without the constraints, proving that the proposed metric learning framework can learn discriminative transformations. Furthermore, we also observe that MDSSL$^{(2)}$ outperforms MDSSL$^{(1)}$ and MDSSL$^{(0)}$ in all cases, so the proportion of MDSSL$^{(2)}$ should be larger than MDSSL$^{(1)}$ and MDSSL$^{(0)}$ in the objective function. In λ_1=0.01 and λ_2=0.1, the value of α_0, α_1 and α_2 are 0.2725, 0.1656 and 0.5619, respectively. These results validate that our analysis is reasonable.

Similarly, we show the performance by varying k_1 and k_2 in Fig. 8. It is obvious to know that their values have less impact

on the performance from the figure. MDSSL can achieve the best performance with larger k_1 and k_2, i.e., $k_1 \in [30, 50]$ and $k_2 \in [300, 500]$.

H. Convergency and Computational Time

In this part, we compare the computational complexity of different methods on the Wiki dataset. In the experiment, all the 1500 paired training samples and 866 paired testing samples are used for evaluating the computational time. Our hardware configuration comprises a 3.6-GHz CPU and a 16GB RAM. Table VI shows the time spent on the training and testing by all methods with the Matlab R2013a software.

We can see that the training time of our approach (optimization time) is larger than the other compared methods' except for Bi-CMSRM. That is because MDSSL computes multi-order statistical features of images and texts to learn the common subspace, which requires more algebraic operation than other methods and hence leads to a higher computational complexity. The time for computing kernelized feature is 593.31 seconds. As for the test time, the proposed method needs about 10.46 seconds for processing 866 paired testing samples, which is just slower than LCFS. This is possibly due to the fact that LCFS uses trace norm constraint leading to sparse transformations. Note that the training is done offline and only once. Thus the training time cost is not as important as that of the testing time.

As for the convergence rate, we show the convergence curve of MDSSL in Fig. 9. We report the value of objective function versus different number of iterations on the Wiki dataset. We can see that MDSSL can achieve the stable performance after about 10 iterations, which is very fast in practice.

Finally, we show some visual examples of MDSSL's retrieved results on two retrieval directions in Fig. 10. Based on intuitive judgement, we can get the observation that MDSSL finds the most similar matchings at a semantic level, i.e., the correlation defined by class labels.

VI. Conclusion

In this paper, a novel method for cross-modal matching problem is proposed and applied to image and text cross-modal retrieval. To enrich the semantic information, the multi-modal data is represented by the multi-order statistical features. Further, the complementary information of multi-order statistics are exploited by integrating multiple metrics among the multi-spaces. Although we restrict the discussion on images and text, the proposed framework is applicable to match the other multimedia. Experiments on two datasets (Wiki and NUS-WIDE) have shown that the proposed method achieves the best performance compared with existing cross-modal methods. In the future, we will investigate on constructing deep structure to better capture the intrinsic semantic relation among heterogeneous data. We will also improve the metric learning framework to better applying in cross-modal retrieval.

References

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

[2] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2407–2414.

[3] R. Socher and F. Li, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 966–973.

[4] X. Chen, A. Hero, and S. Savarese, "Multimodal video indexing and retrieval using directed information," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 3–16, Feb. 2012.

[5] N. Rasiwasia *et al.*, "A new approach to crossmodal multimedia retrieval," in *Proc. 18th ACM Int. Conf. MultiMedia*, 2010, pp. 251–260.

[6] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Pl-ranking: A novel ranking method for cross-modal retrieval," in *Proc. ACM Int. Conf. MultiMedia*, 2016, pp. 1355–1364.

[7] J. Pereira *et al.*, "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.

[8] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4094–4102.

[9] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1059–1074, Jun. 2014.

[10] A. Sharma, A. Kumar, D. Hal, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2160–2167.

[11] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[12] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2014, pp. 823–831.

[13] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1541–1546.

[14] Y. 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.

[15] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. ACM 21st Int. Conf. MultiMedia*, 2013, pp. 877–886.

[16] B. Bai *et al.*, "Learning to rank with (a lot of) word features," *Inf. Retrieval*, vol. 13, no. 3, pp. 291–314, 2010.

[17] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.

[18] Y. Verma and C. Jawahar, "Im2text and text2im: Associating images and texts for cross-modal retrieval," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–13.

[19] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[20] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2088–2095.

[21] Y. Hua, S. Wang, S. Liu, A. Cai, and Q. Huang, "Cross-modal correlation learning by adaptive hierarchical semantic aggregation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1201–1216, Jun. 2016.

[22] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. AAAI 27th Conf. Artif. Intell.*, 2013, pp. 1198–1204.

[23] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Adv. Neural Inform. Process. Syst.*, 2013, pp. 3111–3119.

[25] V. Mahadevan, C. Wong, J. Pereira, T. Liu, N. Vasconcelos, and L. Saul, "Maximum covariance unfolding: Manifold learning for bimodal data," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2011, pp. 918–926.

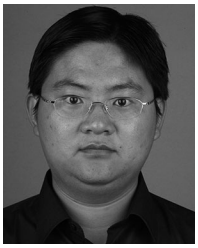
[26] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval," in *Proc. 21st ACM Int. Conf. MultiMedia*, 2013, pp. 897–906.

[27] R. Herbrich, T. Graepel, and T. Obermayer, "Large margin rank boundaries for ordinal regression," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 1999, pp. 115–132.

[28] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, no. 2, pp. 1453–1484, 2006.
[29] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[30] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.
[31] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Int. Conf. Adv. Neural Inform. Process. Syst.*, 2012, pp. 2231–2239.
[32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. MultiMedia*, 2014, pp. 675–678.
[33] X. Li, W. Hu, Z. Zhang, X. Zhang, M. Zhu, and J. Chen, "Visual tracking via incremental log-euclidean Riemannian subspace learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.
[34] D. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[35] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 329–336.
[36] R. Wang, H. Guo, L. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2496–2503.
[37] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian metric for point-to-set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1677–1684.
[38] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, 2007.
[39] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 2015.
[40] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 368–375.
[41] N. Rasiwasia, P. Moreno, and N. Vascondelos, "Bridging the gap: query by semantic example," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 923–938, Aug. 2007.
[42] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1886–1894.

Liang Zhang received the M.S. degree in technology of computer application from the University of Jinan, Jinan Shi, China, in 2014, and is currently working toward the Ph.D. degree at the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China.

His research interests include image and text retrieval, metric learning, and deep learning.



Bingpeng Ma received the B.S. degree in mechanics in 1998 and the M.S. degree in mathematics in 2003, both from the Huazhong University of Science and Technology, Wuhan Shi, China, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009.

He was a Postdoctorial Researcher with the University of Caen, Caen, France, from 2011 to 2012. In March 2013, he joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning. He especially focuses on face recognition, person reidentification, and the related research topics.

Guorong Li received the B.S. degree in technology of computer application from Renmin University of China, Haidian Qu, China, in 2006, and the Ph.D. degree in technology of computer application from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2012.

She is currently an Associate Professor within the University of Chinese Academy of Sciences, Beijing, China. Her research interests include object tracking, video analysis, pattern recognition, and cross-media analysis.

Qingming Huang received the B.S. degree in computer science and the Ph.D. degree in computer engineering, both from the Harbin Institute of Technology, Harbin, China, in 1998 and 1994, respectively.

He is currently a Professor with the University of Chinese Academy of Sciences, Beijing, China, and an Adjunct Research Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He served as a Postdoctoral Fellow with the National University of Singapore, Singapore, from 1995 to 1996, and served as a Member and research staff with the Institute for Infocomm Research, Singapore, from 1996 to 2002. He joined the University of the Chinese Academy of Sciences as a Professor under the Science100 Talent Plan in 2003, and has been granted by the China National Funds for Distinguished Young Scientists in 2010. He also received the National Hundreds and Thousands Talents Project in 2014. He has authored or coauthored more than 300 academic papers in prestigious international journals including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, and VLDB. His research interests include multimedia video analysis, image processing, computer vision, and pattern recognition.

Prof. Huang is an Associate Editor of Acta Automatica Sinica, and a Reviewer of various international journals including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He has served as Program Chair, Track Chair, and the TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, and PSIVT.

Qi Tian received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the Ph.D. degree in ECE from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2002, and the M.S. degree in ECE from Drexel University, Philadelphia, PA, USA in 1996.

He is currently a Full Professor with the Department of Computer Science, University of Texas at San Antonio (UTSA), San Antonio, TX, USA. He was a tenured Associate Professor from 2008 to 2012 and a Tenure-Track Assistant Professor from 2002 to 2008. During 2008 and 2009, he took one-year Faculty Leave with Microsoft Research Asia, Beijing, China, as a Lead Researcher with the Media Computing Group. He has authored or coauthored more than 340 refereed journal and conference papers. His research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian is the Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Multimedia System Journal* (MMSJ), and is on the Editorial Board of *Journal of Multimedia* (JMM), and *Journal of Machine Vision and Applications* (MVA). He is the Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and the *Journal of Computer Vision and Image Understanding*. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He was the recipient of the 2014 Research Achievement Award from College of Science, UTSA. He was the recipient of the 2010 ACM Service Award. He was the coauthor of a Best Paper in ACM ICMR 2015, a Best Paper in PCM 2013, a Best Paper in MMM 2013, a Best Paper in ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Best Student Paper in ICASSP 2006, and coauthor of a Best Student Paper Candidate in ICME 2015, and a Best Paper Candidate in PCM 2007.