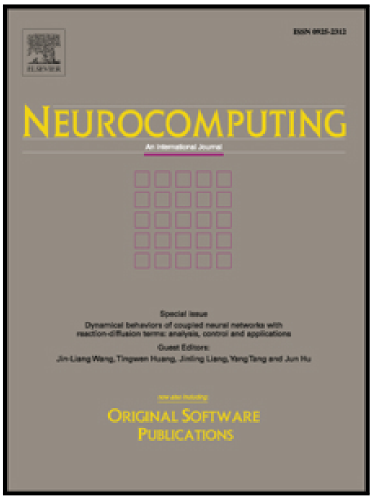


Subspace Learning by Kernel Dependence Maximization for Cross-modal Retrieval

Meixiang Xu, Zhenfeng Zhu, Yao Zhao, Fuming Sun

PII: S0925-2312(18)30530-7
DOI: 10.1016/j.neucom.2018.04.073
Reference: NEUCOM 19549

To appear in: *Neurocomputing*

Received date: 23 November 2017
Revised date: 16 March 2018
Accepted date: 30 April 2018

Subspace Learning by Kernel Dependence Maximization for Cross-modal Retrieval

Meixiang Xu[a,b], Zhenfeng Zhu[a,b,*], Yao Zhao[a,b], Fuming Sun[c]

[a] *Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China*
[b] *Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China*
[c] *School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China*

Abstract

Heterogeneity of multi-modal data is the key challenge for multimedia cross-modal retrieval. To solve this challenge, many approaches have been developed. As the mainstream, subspace learning based approaches focus on learning a latent shared subspace to measure similarities between cross-modal data, and have shown their remarkable performance in practical cross-modal retrieval tasks. However, most of the existing approaches are intrinsically identified with feature dimension reduction on different modalities in a shared subspace, unable to fundamentally resolve the heterogeneity issue well; therefore they often can not obtain satisfactory results as expected. As claimed in Hilbert space theory, different Hilbert spaces with the same dimension are isomorphic. Based on this premise, isomorphic mapping subspaces can be considered as a single space shared by multi-modal data. To this end, we in this paper propose a correlation-based cross-modal subspace learning model via kernel dependence maximization (KDM). Unlike most of the existing correlation-based subspace learning methods, the proposed KDM learns subspace representation for each modality by maximizing the kernel dependence (correlation) instead of directly maximizing the feature correlations between multi-modal data. Specifically, we first map multi-modal data into different Hilbert spaces but with the same dimension individually, then we calculate kernel matrix in each Hilbert space and measure the correlations between multi-

[*] Corresponding author
Email addresses: xumx0721@sina.com (Meixiang Xu), zhfzhu@bjtu.edu.cn (Zhenfeng Zhu), yzhao@bjtu.edu.cn (Yao Zhao), sunwenfriend@hotmail.com (Fuming Sun)

modalities based on kernels. Experimental results have shown the effectiveness and competitiveness of the proposed KDM against the compared classic subspace learning approaches.

Keywords: Correlation/dependence, cross-modal retrieval, subspace learning, supervised learning

2010 MSC: 00-01, 99-00

1. Introduction

Nowadays, multi-modal multimedia data are omnipresent on the Internet and many social websites like YouTube, Facebook, Flickr and so on. In multimedia domain, multi-modal data are referred to as various types of media data such as audio clips, videos, images and texts, etc. While multi-modal data are heterogeneous, they can provide complementary information about the same semantic objects, which is helpful for people to comprehensively understand the semantic objects. Therefore, numerous efforts have been made to study cross-modal retrieval. The goal of cross-modal retrieval is to retrieve the relevant data objects from one modality given one data object from another modality as query.

However, the ‘content-gap’ issue, referring to that the heterogeneity from multi-modal multimedia data makes the similarity between them unable to be directly measured, is still a big challenge for cross-modal retrieval. To address this challenge, two strategies are commonly adopted. One is to directly calculate cross-modal similarities based on the known data relationships by learning cross-modal similarity measures, without obtaining an explicit common subspace that is shared with multi-modal data [18][19][32]. The other is to learn a latent common subspace where cross-modal data similarities can be effectively executed, which is generally termed as common subspace learning. Currently, common subspace learning is one mainstream for cross-modal retrieval. Representative approaches characterized with this type include unsupervised ones [34, 30, 1], supervised ones [28, 33], sparsity-based ones [13, 4], deep neural network based ones [21, 2, 41], etc. Fig. 1 shows the illustration of cross-modal subspace learning.

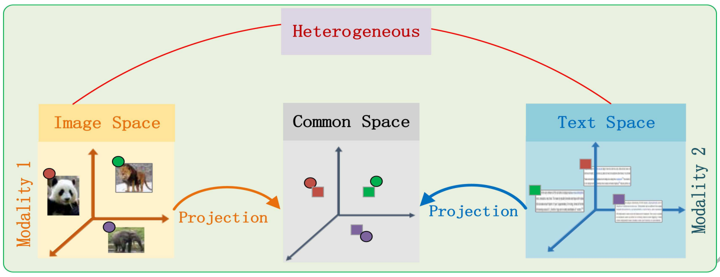


Figure 1: Illustration of Cross-modal subspace learning (Take image and text two modality as an example)

Despite the heterogeneity, multi-modal multimedia data shares with the same semantics since they intrinsically describe the same semantic object. Naturally, there may exist some inherent correlations among them, based on which common subspace representations can be learnt. Accordingly, correlation-based cross-modal subspace learning have been the research focus for cross-modal retrieval. Nevertheless, it is not trivial to directly capture correlations across modalities due to the fact that data from different modalities may have very different statistical properties. As stated in statistics, canonical correlation analysis (CCA) [16], Kullback-Leibler (KL) divergence [6], mutual information [35], Hilbert-Schmidt Independence Criterion (HSIC) [12, 7, 17],and son on, are classical techniques for measuring correlation (dependence) of two random variables. These techniques provide the theoretical foundation for correlation-based learning and have been used for various learning tasks such as dimensionality reduction [35, 47], classification [48, 33], clustering [3], dictionary learning [9, 10], multi-modal retrieval [27], etc. As one of the fundamental statistical correlation analysis approaches, CCA has attracted extensive attention during the past decades. Taking data from two heterogeneous modalities as two group variables, CCA is used to learn a latent common subspace by maximizing the pairwise correlations between them for cross-modal retrieval. Due to its flexibility, a great many of the CCA-like methods are proposed successively [14, 2, 27]. Much more CCA-based variants for common subspace learning can be found in [24].

As the mainstream, common subspace learning approaches aim at learning isomorphic subspace representations for each modality to measure the similarities between cross-modal data, and have shown the great effectiveness to cross-modal retrieval tasks in the last decades, which . However, most of them are intrinsically identified with feature dimension reduction to obtain subspaces with the same dimensions, without specifying whether the subspaces satisfy the isomorphic mapping relationship strictly. Actually, isomorphic mapping subspaces can be considered as the single space shared by multi-modal data, in which the modality similarities are able to be directly measured. To this end, in this paper we seek for such subspace for each modality via kernel dependence maximization based on the Hilbert space theory. Specifically, we learn different mappings which can transform multi- modal data from multiple heterogeneous spaces to multiple isomorphic Hilbert spaces with the same dimensions, individually. Then, we calculate kernel matrix of each modality in each Hilbert space and measure the correlations between the calculated kernel matrices. The experimental results on two benchmark data sets demonstrate the effectiveness and superiority of our proposed model, compared with the other state-of-the-art subspace learning based approaches.

The main contributions of our work can be summarized as the following:

- Unlike most subspace learning based methods that directly project multi-modality data into a latent common subspace to measure similarities between multi-modality samples, the proposed kernel dependence maximization model, to learn subspace representations for cross-modal retrieval, projects multi-modality data into multiple different Hilbert spaces with the same dimensions where cross-modal similarities can be performed across different Hilbert spaces since they are mutually isomorphic.

- The kernel matrix essentially measures the similarities among samples from each modality, therefore intra-modality similarity can be well preserved. Besides, for each modality data, the consistency between feature-based similarity and semantic-based similarity can also be preserved.

- To solve the optimization problem, an efficiently iterative algorithm based on the alternating strategy is designed with its rigorous convergence analysis theoretically, which also has rapid convergence speed within about ten iterations on the tested datasets.

The remainder of the paper is structured as follows. In Section 2, we review some related works on cross-modal subspace learning. In Section 3, some notations used throughout the paper and Hilbert-Schmidt Independence Criteria (HSIC) are introduced. Section 4 provides the proposed cross-modal subspace learning model via kernel dependence maximization approach in details, including the designed optimizing algorithm and convergence analysis with rigorous theoretical proof. The experimental results to show the performance of the proposed model are reported in Section 5. Section 6 concludes the paper.

2. Related Work

In the past decades, there have been many approaches developed for the cross-modal retrieval tasks such as probabilistic models based ones [32], metric learning ones [18, 19], subspace learning based one [15, 37], etc. Thereinto, subspace learning based ones are designed to learn a latent common subspace for directly measuring the similarity between different modalities of data, currently dominating the landscape in cross-modal retrieval. This type of methods can be categorized into three paradigms i.e. unsupervised, semi-supervised and supervised concerning how to exploit the domain knowledge (e.g. pairwise information, label information, etc) in the training phase.

Unsupervised subspace learning based algorithms utilize the correspondence information of one-to-one mapping multi-modal data to learn common subspace representations for each modality, then calculate the pair-wise closeness between modalities. Canonical Correlation Analysis (CCA), Partial Least Squares (PLS) [30] and Bilinear Model (BLM) [34] are the most representative baseline unsupervised ones used for cross-modal retrieval. Although unsupervised ones have been effectively applied to cross-model retrieval, the retrieval performance is not as pleasurable as expected since they do not make use of the discriminant information encoded in the semantic labels. To obtain a more discriminative common subspace where different classes can be well separated as much as possible, supervised methods exploit semantic label information for subspace learning. Taking the semantic label as the third feature view, Gong et al [11] present a three-view CCA (CCA-3V) to learn more discriminative subspace representations for cross-modal retrieval. Extending CCA to perform multi-label cross-modal retrieval, Ranjan et al [27] introduce multi-label CCA (ml-CCA) by incorporating multi-label annotations as high level semantic information. In [26], Semantic Matching (SM) and Semantic Correlation Matching (SCM) are proposed to learn semantic subspace representations for each modality by utilizing logistic regression as the classification way. Simultaneously, SCM has demonstrated the success in combining CCA with semantic label information. Moreover, Generalized Multiview Analysis (GMA) in [31] is another supervised extended version of CCA for cross-modal retrieval. Compared with unsupervised ones, supervised methods can learn a more discriminative subspace owing to the discriminative information implied in the labels. While for semi-supervised methods, part of the training multi-modal data are labeled with semantic class labels, and part are without labels. Typical cross-modal semi-supervised approaches include [37], [44], [25], etc. Specifically, using both labeled and unlabeled data to build the multi-modal graph for preserving two types of similarity relationship i.e. intra-modality and inter-modality, Wang et al propose the Joint Feature Selection and Subspace learning (JFSSL) for cross-modal retrieval [37]. Benefiting from dictionary learning, Xu et al propose a semi-supervised coupled dictionary learning approach jointly using both the paired and unpaired samples to learn sparse subspace representations for the multi-modal data [44]. To execute the matching task, they apply the coupled feature mapping technique in [38] to map the learned sparse representations for each modality into the semantic space. Taking only one graph to jointly model all labeled and unlabeled multi-media data, Peng et al propose the semi-supervised cross-media feature learning algorithm with unified patch graph regularization (S^2UPG) [25].

Beyond that, Deep neural network (DNN), as one popular non-linear relationship learning means, has shown its remarkable ability in various multimedia applications [23, 46, 45] and has been gradually exploited for common subspace learning. Inspired by DNN, Ngiam et al extend restricted Boltzmann machine(RBM) to the bimodal deep autoencoder for common space learning [21]. In the bimodal deep autoencoder, inputs of two different modal data pass through a shared code layer to learn the cross-media correlations and simultaneously to preserve the reconstruction information. Driven by the overwhelming power from deep learning, a variety of deep architectures are developed and perform well in cross-media retrieval [8][29]. Incorporating DNN into CCA, Andrew et al propose the deep canonical correlation analysis (DCCA) [2] for learning the complex non-linear transformations for two media types. In fact, DCCA is a nonlinear extension of CCA. Motivated by CCA and reconstruction-based autoencoders, Wang et al propose deep canonically correlated autoencoders (DCCAE), which is essentially identified with an autoencoder regularized DCCA [41]. To summarize, DNN-based approaches are deep extensions of CCA in neural networks. Though such type of methods have demonstrated their effectiveness, they are often confronted with the difficulty of high computational complexity since tuning parameters involved in neural network requires much efforts in the training phase. Much more algorithms about subspace learning for cross-modal retrieval can refer to the surveys and references therein [39] [24].

This paper is built upon our preliminary conference version [43]. 1) Different from the conference version, we in current version not only exploit the pairwise relationship between multi-modal data but also incorporate the shared semantic label information into subspace learning. By maximizing the dependence between multiple modalities as well as the dependence between each modality and the shared semantic label by measuring kernel similarity jointly, more discriminative subspace representations can be learnt for cross-modal retrieval. Simultaneously, the consistency between feature-based similarity and semantic-based similarity for samples from each modality, and the sample similarity consistency between modalities can be well preserved. 2) We extend the previous objective function for two views to a supervised multi-view version, which can be also used for multi-modal retrieval. 3) We theoretically provide rigorous convergence analysis on the optimizing algorithm.

3. Preliminaries

3.1. Notations

To begin with, we introduce some notations adopted in the paper. For any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{A}^{\cdot\cdot i}$ and $\mathbf{A}^{:j}$ are used to represent its i-th row and j-th column, respectively. $\|\mathbf{A}\|_{2,1}$ is the $\ell_{2,1}$-norm of $\mathbf{A}$, defined as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n} \left\|\mathbf{A}^{\cdot\cdot i}\right\|_2$. $\|\mathbf{A}\|_{HS}$ is the Hilbert-Schmidt norm of $\mathbf{A}$, defined as $\|\mathbf{A}\|_{HS} = \sqrt{\sum_{i,j} a_{ij}^2}$. Besides, $tr\left(\cdot\right)$ represents the trace operator, $\otimes$ the tensor product and $\mathbf{I}$ an identity matrix with an appropriate size. Throughout the paper, matrices and vectors are represented in uppercase and lowercase letters respectively, and both are highlighted in bold. Variables are represented by italic lowercase letters.

3.2. Hilbert-Schmidt Independence Criteria

Let C_{xy} be the cross-covariance function between x and y, $\varphi(x)$ and $\phi(y)$ two mapping functions with $\varphi(x) : x \in \mathcal{X} \rightarrow \mathbb{R}$ and $\phi(y) : y \in \mathcal{Y} \rightarrow \mathbb{R}$, $\mathcal{G}$ and $\mathcal{H}$ two RKHSs in $\mathcal{X}$ and $\mathcal{Y}$. The associated positive definite kernels k_x and k_y is defined as $k_x(x, x^T) =< \Phi(x), \Phi(x) >_{\mathcal{G}}$ and $k_y(y, y^T) =< \Phi(y), \Phi(y) >_{\mathcal{H}}$. Then cross-covariance C_{xy} is defined as:

$$C_{xy} = E_{xy}\left[\left(\varphi\left(x\right) - u_x\right) \otimes \left(\phi\left(y\right) - u_y\right)\right] \tag{1}$$

where u_x and u_y is the expectation of $\varphi(x)$ and $\phi(y)$ respectively, i.e. $u_x = E\left(\varphi\left(x\right)\right)$ and $u_y = E\left(\phi\left(y\right)\right)$.

Given two independent RKHSs $\mathcal{G}$, $\mathcal{H}$ and the joint distribution p_{xy}, HSIC is the Hilbert-Schmidt norm of C_{xy}, defined as:

$$HSIC\left(p_{xy}, \mathcal{G}, \mathcal{H}\right) := \left\|C_{xy}\right\|_{HS}^2 \tag{2}$$

In practical applications, an empirical estimate formulation of HSIC is commonly used. Given N finite number of data samples $Z := \{(x_1, y_1), \cdots, (x_N, y_N)\}$, the empirical expression of HSIC is formulated as:

$$HSIC\left(Z, F, G\right) = \left(n-1\right)^{-2} tr\left(\mathbf{K}_1\mathbf{H}\mathbf{K}_2\mathbf{H}\right) \tag{3}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are two Gram matrices with $k_{1,ij} = k_1(x_i, x_j)$ and $k_{2,ij} = k_2(y_i, y_j)$ $(i, j = 1, \cdots, n)$. $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, is a centering matrix, and $\mathbf{1}_n \in \mathbb{R}^n$ is a full-one column vector.

More details about HSIC can refer to literatures [7, 17].

4. Cross-modal Subspace Learning Model by Kernel Dependence Maximization

4.1. Formulation

Assume that there are M types of multi-modal media training data with n samples, denoted as $\mathbf{X}_v = [\mathbf{x}_1^i, \cdots, \mathbf{x}_v^i] \in \mathbb{R}^{n \times d_v}(i = 1, \cdots, n; v = 1, \cdots, M)$, where d_v is the dimension of each modality and $\mathbf{x}_v^i \in \mathbb{R}^{d_v \times 1}$. The shared semantic label matrix $\mathbf{Y} \in \mathbb{R}^{n \times c}$ is represented as $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^T$, where $\mathbf{y}_i \in \mathbb{R}^{c \times 1}$ and c is the number of the possible semantic labels. If $\mathbf{x}_v^i$ belongs to the j-th $(j = 1, \cdots, c)$ class, we have $y_{ij} = 1$, otherwise $y_{ij} = 0$, where y_{ij} is the j-th element of $\mathbf{y}_i$. The goal of this paper is to learn isometric representations for heterogeneous multi-modal data by maximizing kernel dependence instead of feature correlations of each modality directly. As a kernel itself is in essence a similarity function, each kernel matrix measures the similarity relationship among samples in each modality, which is called the intra-modality similarity. Therefore, our proposed model can simultaneously preserve the intra-modality similarity relationship.

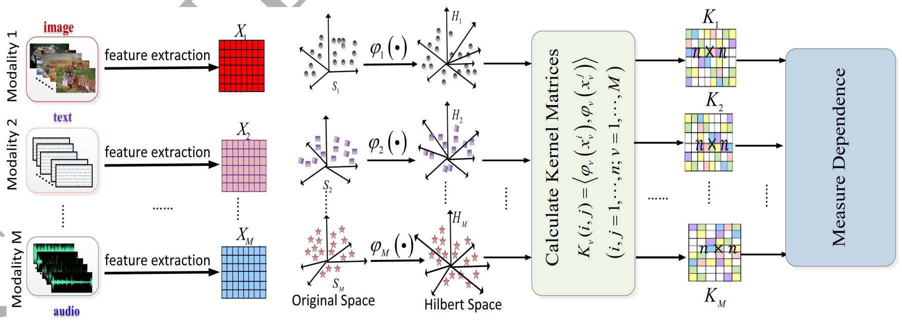


Figure 2: Framework of the proposed model

4.2. Model

First, M heterogeneous multi-modal data are mapped into M different Hilbert spaces H_v under the mapping functions $\varphi_v(\cdot)$, $v = 1, \cdots, M$. According to Hilbert space theory, different Hilbert spaces with the same dimensions are isomorphic, by which we can perform cross-modal retrieval. The new representation $\mathbf{z}_v^i$ for each sample $\mathbf{x}_v^i (i = 1, \cdots, n)$ from modality in H_v is $\mathbf{z}_v^i = \varphi_v(\mathbf{x}_v^i)$. In each Hilbert space, we calculate the kernel matrix $\mathbf{K}_v$ by $\mathbf{K}_v(i, j) = \left\langle \mathbf{z}_v^i, \mathbf{z}_v^j \right\rangle$, $i, j = 1, \cdots, n$. Then we measure the correlations between multi-modal data based on the kernel matrix by similarity measures. The framework of the model is shown in Fig. 2.

For simplicity, we adopt linear kernel as the kernel measure, kernel matrices of multi-modal data K_{X_v} are denoted as $\mathbf{K}_{X_v} = \langle \mathbf{Z}_v, \mathbf{Z}_v \rangle = \mathbf{Z}_v \mathbf{Z}_v^T$, where $\mathbf{z}_v^i = \varphi_v\left(\mathbf{x}_v^i\right) = \mathbf{x}_v^i \mathbf{P}_v \in R^{d_v \times d} (v = 1, \cdots, M)$, and $\mathbf{P}_v$ are the projection matrices. The kernel matrix of the shared semantic label is denoted as $\mathbf{K}_Y = \langle \mathbf{Y}, \mathbf{Y} \rangle = \mathbf{Y}\mathbf{Y}^T$. The similarity between two kernel matrices can be calculated using metric measures such as kernel alignment, Euclidean distance, Kullback-Leibler (KL) divergence, etc.. Here, we adopt the Hilbert-Schmidt Independence Criteria (HSIC). The formulation of the proposed model is:

$$\begin{aligned} &\max_{\mathbf{P}_v} \sum_{\substack{u=1,v=1 \\ u \neq v}}^{M} tr\left(\mathbf{H}\mathbf{K}_{X_u}\mathbf{H}\mathbf{K}_{X_v}\right) + \sum_{v=1}^{M} tr\left(\mathbf{H}\mathbf{K}_{X_v}\mathbf{H}\mathbf{K}_{Y}\right) - \sum_{v=1}^{M} \lambda_v \|\mathbf{P}_v\|_{2,1} \\ &s.t.\ \mathbf{P}_v^T \mathbf{P}_v = \mathbf{I}, \end{aligned} \tag{4}$$

where the first term is to measure the dependence between multiple modality data. The second term is to measure the dependence between each modality and the shared semantic label, which can also preserve the consistency between feature-similarity and semantic-similarity for each sample from each modality. The constraint on $\mathbf{P}_v(v = 1, \cdots, M)$ plays the role of removing redundancy or irrelevant features of the original data. As demonstrated in literatures [22, 40], $\ell_{2,1}$-based learning models have capabilities of sparsity, feature selection and robustness to noise. Therefore, we impose the $\ell_{2,1}$-norm constraint on the projection matrix to expect more discriminative subspace representations for each modality by removing the possible redundant and noisy features contained in the high-dimensional modality data. $\lambda_v(\lambda_v > 0)$ is the regularization parameter.

Without loss of generality, in the following we mainly consider two types of multi-modal multi-media data e.g. image and text. Specifically, the formulation in Eq.(4) is reduced to:

$$\begin{aligned}&\max_{\mathbf{P}_1,\mathbf{P}_2} tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right)+tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{Y}\right)+tr\left(\mathbf{H}\mathbf{K}_{X_2}\mathbf{H}K_{Y}\right)-\lambda_1\|\mathbf{P}_1\|_{2,1}\\&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad-\lambda_2\|\mathbf{P}_2\|_{2,1}\\&s.t.\mathbf{P}_1^T\mathbf{P}_1=\mathbf{I},\mathbf{P}_2^T\mathbf{P}_2=\mathbf{I},\end{aligned}\tag{5}$$

where $\mathbf{K}_{X_1}=\langle\mathbf{Z}_1,\mathbf{Z}_1\rangle=\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T$ and $\mathbf{K}_{X_2}=\langle\mathbf{Z}_2,\mathbf{Z}_2\rangle=\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T$.

4.3. Optimization

By virtue of $\|\mathbf{A}\|_{2,1}=tr\left(\mathbf{A}^T\mathbf{D}\mathbf{A}\right)$, where $\mathbf{D}=diag\left(\frac{1}{\|\mathbf{A}^{\cdot\cdot i}\|_2}\right)$, Eq.(5) can be rewritten as:

$$\begin{aligned}&\max_{\mathbf{P}_1,P_2} tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right)+tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{Y}\right)\\&+tr\left(\mathbf{H}\mathbf{K}_{X_2}\mathbf{H}\mathbf{K}_{Y}\right)-\lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right)-\lambda_2 tr\left(\mathbf{P}_2^T\mathbf{D}_2\mathbf{P}_2\right)\\&s.t.\mathbf{P}_1^T\mathbf{P}_1=\mathbf{I},\mathbf{P}_2^T\mathbf{P}_2=\mathbf{I},\end{aligned}\tag{6}$$

where $\mathbf{D}_v=diag\left(\frac{1}{\|\mathbf{P}_v^{\cdot\cdot i}\|_2}\right)$ and $\mathbf{P}_v^{\cdot\cdot i}$ is the i-th row of $\mathbf{P}_v(v=1,2)$.

To optimize the objective function in Eq.(6) (or Eq.(5)), we employ the alternative optimization strategy. Specifically, according to the alternative optimization rules, the original optimization problem can be decomposed into the following two sub-maximization ones:

4.3.1. Solve $\mathbf{P}_1$ by fixing $\mathbf{P}_2$

Fixing $\mathbf{P}_2$, we can obtain the following equivalent optimization problem of Eq.(6):

$$\begin{aligned}&\max_{\mathbf{P}_1} tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right)+tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{Y}\right)-\lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right)\\\Leftrightarrow&\max_{\mathbf{P}_1} tr\left(\mathbf{P}_1^T\mathbf{B}_1\mathbf{P}_1\right)\\&s.t.\mathbf{P}_1^T\mathbf{P}_1=\mathbf{I},\end{aligned}\tag{7}$$

where $\mathbf{B}_1 = \mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2\mathbf{P}_2^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1 + \mathbf{X}_1^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{X}_1 - \lambda_1\mathbf{D}_1$. Performing eigenvalue decomposition on $\mathbf{B}_1$, we can obtain $\mathbf{P}_1$, which here consists of the first d eigenvectors corresponding to the d largest eigenvalues of $\mathbf{B}_1$.

4.3.2. Solve $\mathbf{P}_2$ by fixing $\mathbf{P}_1$

Likewise, fixing $\mathbf{P}_1$ we can obtain the following equivalent optimization problem:

$$\begin{aligned}
&\max_{\mathbf{P}_2} tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right) + tr\left(\mathbf{H}\mathbf{K}_{\mathbf{X}_2}\mathbf{H}\mathbf{K}_{Y}\right) - \lambda_2 tr\left(\mathbf{P}_2^T\mathbf{D}_2\mathbf{P}_2\right) \\
\Leftrightarrow &\max_{\mathbf{P}_2} tr\left(\mathbf{P}_2^T\mathbf{B}_2\mathbf{P}_2\right) \\
&s.t. \mathbf{P}_2^T\mathbf{P}_2 = \mathbf{I},
\end{aligned} \tag{8}$$

where $\mathbf{B}_2 = \mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1\mathbf{P}_1^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2 + \mathbf{X}_2^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{X}_2 - \lambda_2\mathbf{D}_2$. Likewise, performing eigenvalue decomposition on $\mathbf{B}_2$, we can obtain $\mathbf{P}_2$, consisting of the first d eigenvectors corresponding to the d largest eigenvalues of $\mathbf{B}_2$.

To better understand the procedure for solving the proposed model, we summarize in detail the solver for solving the optimization problem in Eq.(5) as Algorithm 1.

4.4. Convergence Analysis

The convergence of the proposed KDM under the iterative optimization algorithm in Algorithm 1 can be summarized by the following Theorem 1.

Theorem 1. *Under the iterative optimizing rules in Algorithm 1, the objective function defined by Eq.(5) is increasing monotonically, and finally it converges to the global maximum.*

To prove Theorem 1, we need to introduce the following lemma 1. For details about Lemma 1, please refer to the literature [22].

Lemma 1. *For any nonzero $\mathbf{f} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^n$, the following equality holds:*

$$\|\mathbf{f}\|_2 - \frac{\|\mathbf{f}\|_2^2}{2\|\mathbf{g}\|_2} \leq \|\mathbf{g}\|_2 - \frac{\|\mathbf{g}\|_2^2}{2\|\mathbf{g}\|_2} \tag{9}$$

Resorting to the above Lemma 1, the detailed proof of Theorem 1 is provided below.

Algorithm 1 : Cross-modal Subspace Learning via Kernel Dependence Maximization (KDM)

```
Input:
    Multi-modal data $\mathbf{X}_v \in \mathbb{R}^{n \times d_v}$, $v = 1, 2$; the regularization parameters $\lambda_1$ and $\lambda_2$;
Output:
    The projection matrices $\mathbf{P}_v$, $v = 1, 2$;
Initializing: Initialize $\mathbf{P}_1$ and $\mathbf{P}_2$ randomly, and $t = 0$;
while the objective function not converge do
    Update $\mathbf{D}_1$ and $\mathbf{D}_2$: $\mathbf{D}_1^{(t)} = diag\left(\frac{1}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}\right)$, $\mathbf{D}_2^{(t)} = diag\left(\frac{1}{2\left\|\mathbf{P}_2^{\cdot\cdot i(t)}\right\|_2}\right)$;
    Update $\mathbf{P}_1$: obtain $\mathbf{P}_1^{(t+1)}$ by performing eigen-decomposition on $\mathbf{B}_1 = \mathbf{X}_1^T\mathbf{H}\mathbf{X}_2\mathbf{P}_2^{(t)}\left(\mathbf{P}_2^{(t)}\right)^T\mathbf{X}_2^T\mathbf{H}\mathbf{X}_1 + \mathbf{X}_1^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{X}_1 - \lambda_1\mathbf{D}_1^{(t)}$. The $d$ eigenvectors corresponding to the first largest $d$ eigenvalues of $\mathbf{B}_1$ compose $\mathbf{P}_1$ ;
    Update $\mathbf{P}_2$: obtain $\mathbf{P}_2^{(t+1)}$ by performing eigen-decomposition on $\mathbf{B}_2 = \mathbf{X}_2^T\mathbf{H}\mathbf{X}_1\mathbf{P}_1^{(t)}\left(\mathbf{P}_1^{(t)}\right)^T\mathbf{X}_1^T\mathbf{H}\mathbf{X}_2 + \mathbf{X}_2^T\mathbf{H}\mathbf{Y}\mathbf{Y}^T\mathbf{H}\mathbf{X}_2 - \lambda_2\mathbf{D}_2^{(t)}$. The $d$ eigenvectors corresponding to the first largest $d$ eigenvalues of $\mathbf{B}_2$ compose $\mathbf{P}_2$;
    $t = t + 1$;
end while
```

Proof. To prove Theorem 1 can be divided into two processes i.e. (a) Solve $\mathbf{P}_1$ by fixing $\mathbf{P}_2$; (b) Solve $\mathbf{P}_2$ by fixing $\mathbf{P}_2$. Obviously, these two processes are symmetrical, therefore we only need to prove one of them. Below we will give the detailed proof of process (a).

As shown in the previous section, with $\mathbf{P}_2$ fixed, the obtained $\mathbf{P}_1$ by solving the optimization problem in Eq.(5) is the solution to the following problem:

$$\begin{aligned} &\max_{\mathbf{P}_1} tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right) + tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{Y}\right) - \lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right) \\ &s.t. \mathbf{P}_1^T\mathbf{P}_1 = \mathbf{I}. \end{aligned} \tag{10}$$

Define $J(\mathbf{P}_1) = tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{X_2}\right) + tr\left(\mathbf{H}\mathbf{K}_{X_1}\mathbf{H}\mathbf{K}_{Y}\right)$ and let O denote the objective function in Eq.(10), then $O(\mathbf{P}_1) = J\left(\mathbf{P}_1\right) - \lambda_1 tr\left(\mathbf{P}_1^T\mathbf{D}_1\mathbf{P}_1\right)$. In the t-th iteration, there goes:

$$
\begin{aligned}
&\mathbf{P}_1^{(t+1)} = \arg\max\left(J\left(\mathbf{P}_1\right) - \lambda_1 tr\left(\mathbf{P}_1^T \mathbf{D}_1^{(t)} \mathbf{P}_1\right)\right)\\
&\Rightarrow J\left(\mathbf{P}_1^{(t+1)}\right) - \lambda_1 tr\left(\left(\mathbf{P}_1^{(t+1)}\right)^T \mathbf{D}_1^{(t)} \mathbf{P}_1^{(t+1)}\right) \geq J\left(\mathbf{P}_1^{(t)}\right) - \lambda_1 tr\left(\left(\mathbf{P}_1^{(t)}\right)^T \mathbf{D}_1^{(t)} \mathbf{P}_1^{(t)}\right)\\
&\Rightarrow J\left(\mathbf{P}_1^{(t+1)}\right) - \lambda_1 \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t+1)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2} \geq J\left(\mathbf{P}_1^{(t)}\right) - \lambda_1 \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}\\
&\Rightarrow J\left(\mathbf{P}_1^{(t+1)}\right) - \lambda_1 \left\|\mathbf{P}_1^{(t+1)}\right\|_{2,1} + \lambda_1 \left(\left\|\mathbf{P}_1^{(t+1)}\right\|_{2,1} - \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t+1)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}\right)\\
&\geq J\left(\mathbf{P}_1^{(t)}\right) - \lambda_1 \left\|\mathbf{P}_1^{(t)}\right\|_{2,1} + \lambda_1 \left(\left\|\mathbf{P}_1^{(t)}\right\|_{2,1} - \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}\right)
\end{aligned}
\tag{11}
$$

According to Lemma 1, we have:

$$
\left\|\mathbf{P}_1^{(t+1)}\right\|_{2,1} - \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t+1)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2} \leq \left\|\mathbf{P}_1^{(t)}\right\|_{2,1} - \sum_{i=1}^{d} \frac{\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2^2}{2\left\|\mathbf{P}_1^{\cdot\cdot i(t)}\right\|_2}
\tag{12}
$$

Combining Eq.(11) with Eq.(12), we can obtain:

$$
\begin{aligned}
&J\left(\mathbf{P}_1^{(t+1)}\right) - \lambda_1 \left\|\mathbf{P}_1^{(t+1)}\right\|_{2,1} \geq J\left(\mathbf{P}_1^{(t)}\right) - \lambda_1 \left\|\mathbf{P}_1^{(t)}\right\|_{2,1}\\
&\Rightarrow O\left(\mathbf{P}_1^{(t+1)}\right) \geq O\left(\mathbf{P}_1^{(t)}\right)
\end{aligned}
\tag{13}
$$

Likewise, when updating $\mathbf{P}_2$ with $\mathbf{P}_1$ fixed, we can also prove that $O\left(\mathbf{P}_2^{(t+1)}\right) \geq O\left(\mathbf{P}_2^{(t)}\right)$. Further, we can reach $O\left(\mathbf{P}_1^{(t+1)}, \mathbf{P}_2^{(t+1)}\right) \geq O\left(\mathbf{P}_1^{(t)}, \mathbf{P}_2^{(t)}\right)$.

From all the above, the objective function in Eq.(5) under the iterative updating rules designed in Algorithm 1 is non-decreasing. Meanwhile, it is noted that the optimization problem itself is convex. According to the literature [36], the objective function will finally converge to its global optimal solution.

This completes the proof of Theorem 1. □

4.5. Computational Complexity Analysis

In this subsection, we will briefly analyze the computational complexity of the proposed KDM. The complexity of the optimizing algorithm mainly comes from calculating eigenvalue decomposition on $\mathbf{B}_1$ and $\mathbf{B}_2$ to obtain $\mathbf{P}_1$ and $\mathbf{P}_2$. In each iteration, the cost for eigen-decomposition is $o(\sum_{m=1}^{M} d_m^3)$(here $M = 2$). Assume that the the objective function of KDM converges after t_o iterations, the overall cost for KDM is roughly $o(t_o \sum_{m=1}^{M} d_m^3)$.

5. Experiments

To test the performance of the proposed KDM for cross-modal retrieval, experiments were conducted on two real-world datasets. Given a cross-modal problem, using the iterative algorithm in Algorithm 1, we can learn projection matrices on the training set. After that, data from different modalities can be projected into a common subspace, where we can measure the relevance of projected data from each modality. In the testing phase, taking data in one modality as a query set, we can retrieve the relevant data from another modality. To measure the similarity, we adopt the normalized correlation (NC), which shows the best performance for cross-modal retrieval as demonstrated in [26].

5.1. Datasets

Two benchmark datasets, i.e. Wikipedia [42] and NUS-WIDE [5], are used in our experiment. Table 1 gives the statistical information of them and the follow-ups are brief descriptions on them.

- **Wikipedia:** This dataset consists of 2866 image-text pairs that are labeled with 10 semantic classes in total. For each image-text pair, we extract 4096-dimensional visual features by convolutional neural network to represent the image view, and 100-dimensional LDA textual features to represent the text view. In the experiment, the dataset is partitioned into two parts, one for training (2173 pairs) and the other for testing (693 pairs).

- **NUS-WIDE:** This dataset is a subset from [5], including 190420 image examples totally, each with 21 possible labels. For each image-text pair, we extract 500-dimensional SIFT BoVW features for image and 1000-dimensional text annotations for text. To reduce the computational complexity, further we sample a subset with 8687 pairs of image-text. Likewise, the dataset is divided into two parts, one for training (5212 pairs) and the other for testing (3475 pairs).

Table 1: Statistics of Datasets

Datasets	d_1	d_2	$\#labels$	$\#$total	$\#$training	$\#$testing
Wikipedia	4096	100	10	2866	1173	693
NUS-WIDE	500	1000	21	8687	5212	3475

5.2. Evaluation Metric

In the field of information retrieval, Precision, Recall and Mean Average Precision (MAP) are three commonly used evaluation metrics [28]. In our experiments, we mainly adopt MAP to evaluate the retrieval performance. Given a set of queries, MAP refers to the average precision (AP) of all queries, while the average precision (AP) of each query is defined as:

$$AP = \frac{1}{R}\sum_{k=1}^{n_{te}} \frac{R_k}{k} \times rel_k, \tag{14}$$

where n_{te} ist the total number of the testing set, R is the number of relevant data in the returned query items, R_k is the number of relevant data in the first returned k query items. If the item at rank k is relevant, $rel(k) = 1$, otherwise $rel(k) = 0$.

5.3. Benchmark Approaches and Experimental Setup

The proposed KDM for cross-modal subspace learning is supervised, kernel-based and correlation-based. Due to the sparse property of $\ell_{2,1}$-norm, it is sparsity-based in a sense. Consequently, we compare with the following approaches including unsupervised ones i.e. CCA, KPCA [20], KCCA [1] and SCCA [13], together with supervised ones i.e. CCA-3V [11], ml-CCA [27] and LCFS [38]. We also compare with the previous version of our method UDM [43]. Besides, Table 2 makes a summary about all the compared approaches, where 'Y' means YES, indicating one method belongs to the corresponding ascription and 'N' means NO, indicating one method does not belong to the corresponding ascription. It should be noted that, the parameters involved in the compared approaches are set to their default values as in the corresponding literatures. In the following, we will present the specific settings for the parameters involved in KDM.

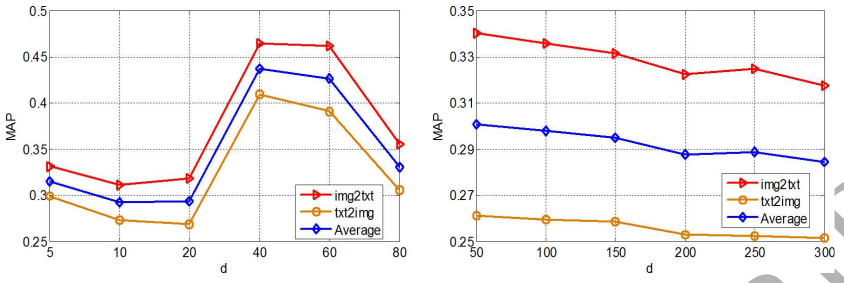
Table 2: Summarization of Compared Algorithms

Methods	unsupervised	supervised	kernel-based	correlation-based	sparsity-based
CCA	Y	N	N	Y	N
KPCA	Y	N	Y	N	N
KCCA	Y	N	Y	Y	N
SCCA	Y	N	N	Y	Y
CCA-3V	N	Y	N	Y	N
ml-CCA	N	Y	N	Y	N
LCFS	N	Y	N	N	N
UDM	Y	N	Y	Y	Y
KDM	N	Y	Y	Y	Y

It can be observed that the proposed KDM mainly involve three parameters, two explicit regularization ones λ_1 and λ_2, as well as an implicit one d, where d is the dimension of the shared subspace. To guarantee that KDM can achieve its optimal performance, by setting different values for the three parameters we perform tests on two datasets. First we fix λ_1 and λ_2 to decide the optimal d. Specifically, we tune d from the range of $\{5, 10, 20, 40, 60, 80\}$ and $\{50, 100, 150, 200, 250, 300\}$ on Wikipedia and NUS-WIDE, respectively. Fig. 3 displays MAP scores of cross-modal retrieval versus different d values on two datasets. As can be seen from Fig. 3, when d takes the value of 40 and 50 on Wikipedia and NUS-WIDE respectively, KDM can obtain its best performance. Consequently, in the following experiments, we set $d = 40$ and $d = 50$ for Wikipedia and NUS-WIDE. Afterwards, we fix d to determine the optimal value of λ_1 and λ_2 by tuning them from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5\}$. Empirically, we determine $\lambda_1 = 10^{-3}$ and $\lambda_2 = 10^3$, as well as $\lambda_1 = \lambda_2 = 10^{-5}$ on Wikipedia and NUS-WIDE, respectively. In the subsequent section, we will give the parameter sensitivity analysis on λ_1 and λ_2.

5.4. Results

According to the above settings, we test the performance of the proposed KDM on the cross-modal retrieval task. Experimental results in comparison with CCA, kernel-based KPCA and KCCA, Sparse CCA and supervised methods CCA-3V, ml-CCA and



(a) Wikipedia

(b) NUS-WIDE

Figure 3: MAP vs. varying d on Wikipedia and NUS-WIDE by fixing λ_1 and λ_2

LCSF, are displayed in Table 3 and Table 4 on Wikipedia and NUS-WIDE, respectively. From the reported results as in both Table 3 and Table 4, we can observe that the proposed KDM achieves the best performance among all the compared approaches, which may benefit from the most discriminative features by jointly using semantic information, the consistency between two types of similarities for samples of multi-modal data, i.e, feature-based similarity and semantic-based similarity, and the good feature selection from $\ell_{2,1}$-norm characterized with sparsity. Besides, it can be seen that CCA-3V, ml-CCA and LCFS have an advantage over CCA, KCCA, KPCA, and SCCA, which lies in that the first three utilize the semantic label information while the last four not. Overall, supervised subspace learning methods obtain better performance can than unsupervised ones, which illustrates that using semantic prior knowledge can benefit learning more discriminative feature representations for cross-modal subspace learning.

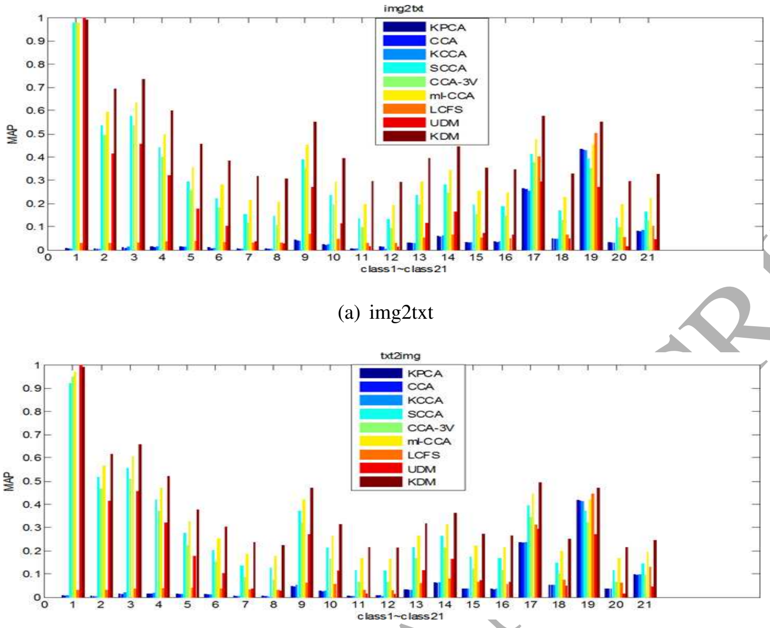
Fig.4 and Fig.5 show the per-class MAP performance of all subspace learning based methods on NUS-WIDE and Wikipedia respectively. As can be seen from the displayed results, compared with unsupervised ones like CCA, KPCA, KCCA, SCCA and UDM, supervised methods like CCA-3V, ml-CCA, LCFS and KDM achieve better results on the per class in most cases, which shows that using label information can facilitate more discriminative subspace representations for cross-modal retrieval. While among the supervised subspace learning approaches, our proposed KDM performs the best, one main reason behind this is that KDM not only incorporates label information

Table 3: MAP Comparison on Wikipedia

Approaches	Image as query	Text as query	Average
KPCA	0.1983	0.1826	0.1905
CCA	0.1222	0.1189	0.1206
KCCA	0.3337	0.3031	0.3184
SCCA	0.2270	0.1961	0.2116
CCA-3V	0.4013	0.3672	0.3843
ml-CCA	0.3634	0.3156	0.3395
LCFS	0.4198	0.3966	0.4082
UDM	0.4204	0.4394	0.4299
KDM	**0.4562**	**0.4785**	**0.4674**

Table 4: MAP Comparison on NUS-WIDE

Approaches	Image as query	Text as query	Average
KPCA	0.2326	0.2215	0.2171
CCA	0.2441	0.2356	0.2399
KCCA	0.2554	0.2451	0.2503
SCCA	0.2415	0.2145	0.2280
CCA-3V	0.3126	0.2757	0.2942
ml-CCA	0.2872	0.2513	0.2693
LCFS	0.3288	0.2674	0.2981
UDM	0.2904	0.2498	0.2702
KDM	**0.3452**	**0.2841**	**0.3147**



(a) img2txt

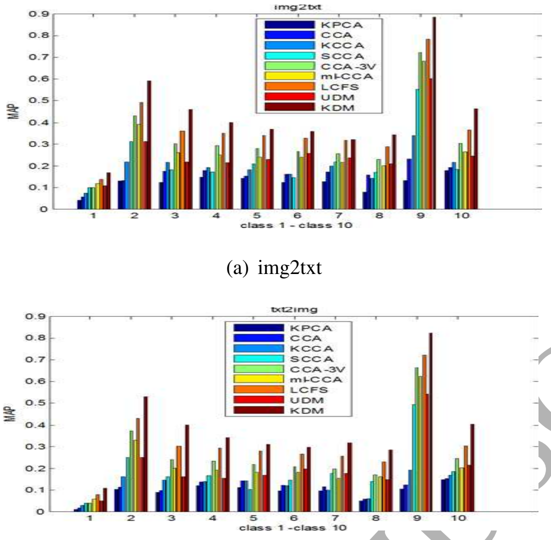
(b) txt2img

Figure 4: Per class MAP on Wikipedia

into subspace learning but also exploit the consistency between feature-similarity and semantic-similarity of each sample from each modality. Another reason is $\ell_{2,1}$-norm constraint on the projection matrix which enables to select more discriminative and label-specific features from the raw features of each modality.

5.5. Parameter Sensitivity Analysis

As formulated in Eq.(5), two parameters λ_1 and λ_2 are involved in the proposed KDM. To show the impacts of them on cross-modal retrieval, we have carried out experiments on Wikipedia and NUS-WIDE by tuning these two parameters from the range of $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5\}$. Experimental results on two datasets for image query vs. text database and text query vs. image database tasks are shown in Fig.6 and Fig.7, respectively. From the two figures, we can observe that the performance of KDM varies as λ_1 and λ_2 change. By contrast, the proposed KDM on Wikipedia is much more sensitive to two parameters than on NUS-



(a) img2txt

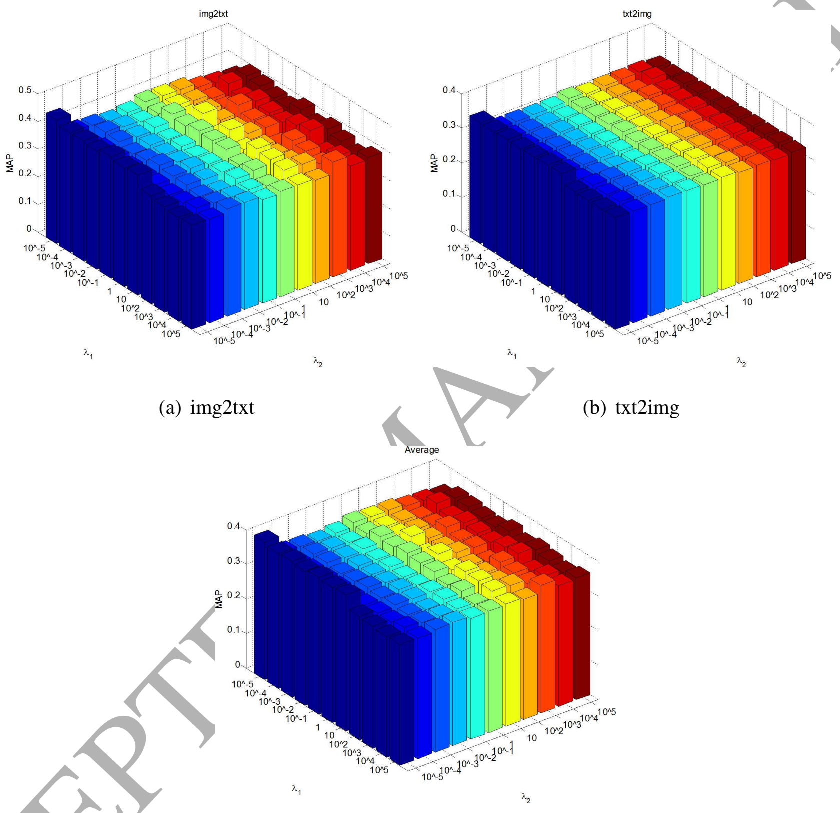
(b) txt2img

Figure 5: Per class MAP on NUS-WIDE

WIDE. While on Wikipedia, the changing trend of MAP in terms of two retrieval tasks is accord with each other. In addition, the proposed KDM on the dataset NUS-WIDE performs better when λ_2 take the value of 10^{-5}.

5.6. Convergence Study

In the preceding part, the convergence behavior of the proposed algorithm has been theoretically analyzed. In the following, we will demonstrate the convergence property experimentally. Specifically, we test convergence in each iteration by computing the values of the objective function in Eq.(5). The convergence criterion used in our test is $\Phi_{t+1} - \Phi_t < 10^{-6}$, where Φ_t is the objective function value in the t-th iteration. Fig.8 displays the relationship between the objective function and the number of iteration on two datasets i.e. Wikipedia and NUS-WIDE. As can be observed from Fig.8, for each dataset, the objective function defined as Eq.(5) can quickly converge to its maximum within about ten iterations, which right verifies the theoretical analysis as discussed before and also shows the efficiency of the designed iterative optimization strategy in Algorithm 1.

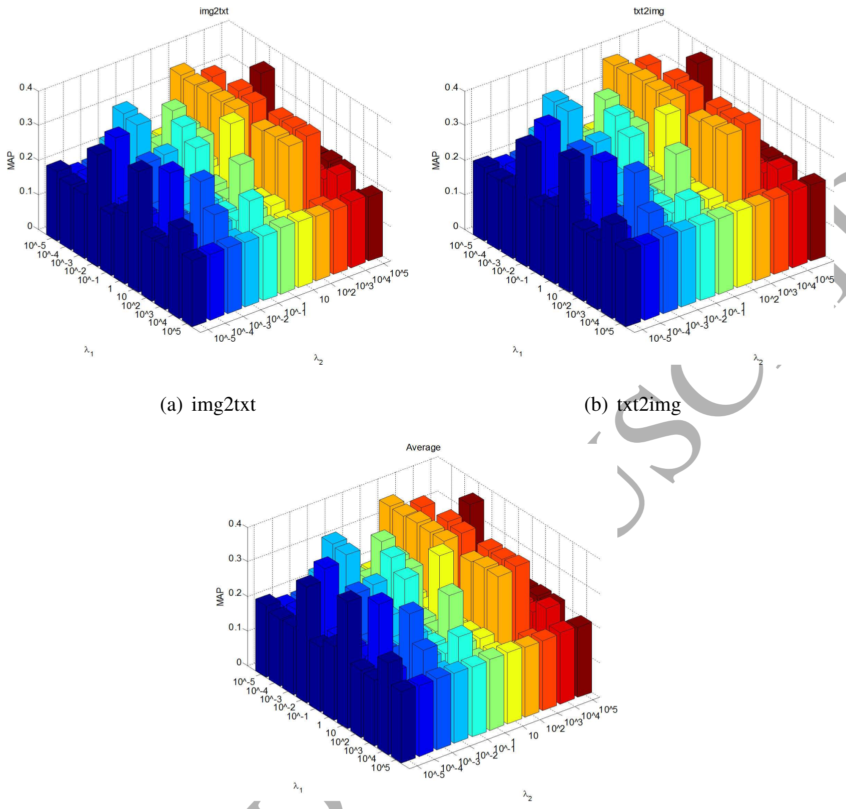


(a) img2txt

(b) txt2img

(c) Average

Figure 6: MAP vs. varying λ_1 and λ_2 on NUS-WIDE

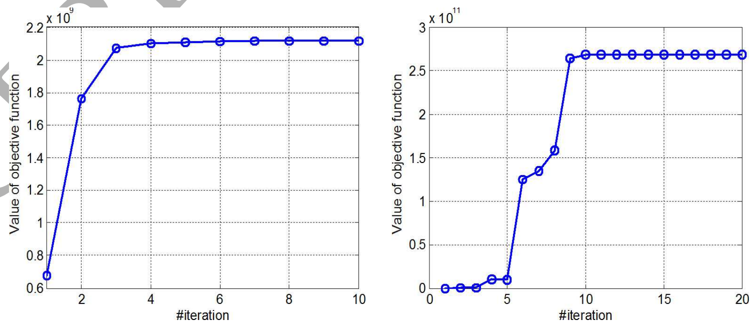


(a) img2txt

(b) txt2img

(c) Average

Figure 7: MAP vs. varying λ_1 and λ_2 on Wikipedia



(a) Wikipedia

(b) NUS-WIDE

Figure 8: The objective function vs. the number of iteration

6. Conclusion

In this paper, we propose the cross-modal subspace learning model via kernel dependence maximization. Unlike most of the existing correlation-based subspace learning methods, the proposed KDM learns subspace representations for each modality by maximizing the kernel dependence instead of directly maximizing the feature correlations between multi-modal data. Moreover, KDM can also preserve the consistency between feature -similarity and semantic-similarity of samples from each modality. Experimental results have demonstrated the effectiveness of the proposed algorithm and show its great competitiveness against the compared classic subspace learning based approaches.

Acknowledgment

The authors must thank the anonymous reviewers for their constructive comments and valuable suggestions on this paper.

References

References

[1] S. Akaho. A kernel method for canonical correlation analysis. In *The International Meeting of the Psychometric Society*, 2007.

[2] G. Andrew, R. Arora, J. Blimes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of International Conference on Machine Learning*.

[3] X.C. Cao, C.Q. Zhang, H.Z. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[4] D.L. Chu, L.Z. Liao, M.K. Ng, and X.W. Zhang. Sparse canonical correlation analysis: new formulation and algorithm. *Transactions on Pattern Analysis and Machine Intelligence*, 35(12):3050–3065, 2013.

[5] T.S. Chua, J.H. Tang, R.C. Hong, H.J. Li, Z.P. Luo, and Y.T. Zhang. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video*, 2009.

[6] A. Cichocki and H.H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 3:757–763, 1996.

[7] J.V. Davis, B.K., P.J.and S. Sra, and I.S. Dhillon. Information-theoretic metric learning. In *Proceedings of International Confernce on Machine Learning*, 2007.

[8] F.X. Feng, R.F. Li, and X.J. Wang. Deep correspondence restricted boltzmann machine for cross-modal retrieval. *Neurocomputing*, 154:50–60, 2015.

[9] M.J. Gangeh, P. Fewzee, A. Ghodsi, and M. Fakhri. Kernelized supervised dictionary learning. *IEEE Transactions on Signal Processing*, 61(19):4753–4767, 2013.

[10] M.J. Gangeh, P. Fewzee, A. Ghodsi, and M. Fakhri. Multi-view supervised dictionary learning in speech emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(6):1056–1068, 2014.

[11] Y.C. Gong, Q.F. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, pages 1–24, 2013.

[12] A. Gretton, O. Bousquet, A. Smola, and B. Schlkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of International Conference on Algorithmic Learning Theory*, 2005.

[13] D.R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[14] D.R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[15] R. He, M. Zhang, L. Wang, Y. Ji, and Q.Y. Yin. Cross-modal subspace learning via pairwise constraints. *IEEE Transaction on Image Processing*, 24(12):5543–5556, 2015.

[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[17] J.C.Principe. Information theory, machine learning, and reproducing kernel hilbert spaces. *Information Science and Statistics*, pages 1–45, 2010.

[18] Y.Q. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision*, 2011.

[19] S.Q. Jiang, X.H. Song, and Q.M. Huang. Relative image similarity learning with contextual information for internet cross-media retrieval. *Multimedia System*, 20(6):645–657, 2014.

[20] S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y.M. Cheng, and H.J. Zhan. Kernel machine based learning for multi-view face detection and pose estimation. In *IEEE International Conference on Computation Vision*.

[21] J.Q. Ngiam, A. Khosla, M.Y. Kim, J.H. Nam, H. Lee, and A.Ng. Multimodal deep learning. In *Proceedings of International Conference on Machine Learning*, 2011.

[22] F.P. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In *Advances in Neural Information Processing Systems*.

[23] Y.X. Peng, X. Huang, and J.W. Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

[24] Y.X. Peng, X. Huang, and Y.Z. Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, PP(99):1–13, 2017.

[25] Y.X. Peng, X.H. Zhai, Y.Z. Zhao, and X. Huang. Semi-supervised cross-media feature learning with unified path graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):583–596, 2016.

[26] J.C. Pereira, G. Doyle, N. Rasiwasia, Gert R. G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[27] V. Ranjan, N. Rasiwasia, and C.V. Jawahar. Multi-label cross-modal retrieval. In *Proceedings of Interantional Conference on Computer Vision*.

[28] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G Lanckriet, and R. Levy ahd N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the ACM International conference on Multimedia*.

[29] J. Shao, L.Q. Wang, Z.C. Zhao, F. Su, and A.N. Cai. Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. *Neurocomputing*, 214:618–628, 2016.

[30] A. Sharma and D.W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, lowresolution and sketch. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[31] A. Sharma, A. Kumar, and H. Daume III. Generalized multi-view analysis: a discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[32] G.L. Song, S.H. Wang, Q.M. Huang, and Q.Tian. Multimodal similarity gaussian process latent variable model. *IEEE Transanctions on Image Prosessing*, 26(9):4168–4181, 2017.

[33] K. Tae-Kyun, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classess using canonical correlation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[34] J.B. Tenenbaum and W.T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[35] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3(3):1415–1438, 2003.

[36] S. Boydand L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[37] K.Y. Wang, R. He, L. Wang, W. Wang, and T.N. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, 38(10):2010–2023, 2016.

[38] K.Y. Wang, R. He, W. Wang, L. Wang, and T.N. Tan. Learning coupled feature spaces for cross-modal matching. In *Proceedings of International Conference on Computer Vision*.

[39] K.Y. Wang, Q.Y. Yin, W. Wang, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv:1607.06215v1[cs.MM]*, 2016.

[40] T.H. Wang, W. Li, and X.W. He. Kernel learning with hilbert-schmidt independence criterion. In *Chinese Conference on Pattern Recognition*.

[41] W.R. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of International Conference on Machine Learning*, 2015.

[42] Y.C. Wei, Y. Zhao, C.Y. Lu, S.K. Wei, L.Q Liu, Z.F. Zhu, and S.C. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Transactions on Cybernetics*, 47(2):449–460, 2017.

[43] M.X. Xu, Z.F. Zhu, and Y. Zhao. Unsupervised multi-view subspace learning via maximizing dependence. In *Chinese Conference on Computer Vision*, 2017.

[44] X. Xu, Y. Yang, A. Schimada, R. Taniguchi, and L. He. Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In *Proceedings of ACM International Conference on Multimedia*, 2015.

[45] C.G. Yan, H.T. Xie, S. Liu, J. Yin, Y.D. Zhang, and Q.H. Dai. Effective uyghur language text detection in complex background images for traffic prompt identification. *IEEE Transanctions on Intelligent Transportation Systems*, 19(1):220–229, 2017.

[46] C.G. Yan, H.T. Xie, D.B. Yang, J. Yin, Y.D. Zhang, and Q.H. Dai. Supervised hash coding with deep neural network for environment perception of intelligent vehicles. *IEEE Transanctions on Intelligent Transportation Systems*, 9(1):284–295, 2017.

[47] Y. Zhang and Z.H. Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 4(3):14, 2010.

[48] S.H. Zhu, X. Ji, W. Xu, and Y.H. Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*.

Meixiang Xu received the M.E. degree from Liaoning University of Technology, Jinzhou, China, in 2015. Currently, she is pursuing the Ph.D. degree in the Institute of Information Science, Beijing Jiaotong University, Beijing, China. Her research interests include data analysis, cross-modal retrieval and machine learning.

Zhenfeng Zhu received the Ph.D. degree in pattern recognition and intelligence system from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005. He was a Visiting Scholar with the Department of Computer Science and Engineering, Arizona State University, AZ, USA, in 2010. He is currently a Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His current research interests include image and video understanding, computer vision, and machine learning.

Yao Zhao received the B.S. degree from Fuzhou University, Fuzhou, China, in 1989, and the M.E. degree from Southeast University, Nanjing, China, in 1992, both from the Radio Engineering Department, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor in 1998 and a Professor in 2001, both with BJTU. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. He is currently leading several national research projects from the 973 Program, 863 Program, and the National Science Foundation of China. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the editorial boards of several international journals, including as an Associate Editor of the IEEE Transactions on Cybernetics, Associate Editor of the IEEE Signal Processing Letters, the Area Editor of Signal Processing: Image Communication, and an Associate Editor of Circuits, System, and Signal Processing. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of the Ministry of Education of China in 2013. He is an IET Fellow and an IEEE senior member.

Fuming Sun received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2007. From 2012 to 2013, he was a Visiting Scholar with the Department of Automation, Tsinghua University. He is currently a Professor with the School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou, China. His current research

interests include content-based image retrieval, image content analysis, and pattern recognition.