

Multi-label double-layer learning for cross-modal retrieval



Jianfeng He^{a,b}, Bingpeng Ma^{a,b,*}, Shuhui Wang^b, Yugui Liu^a, Qingming Huang^{a,b}

^aSchool of Computer and Control Engineering, University of China Academy Science, Beijing, China

^bKey Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:

Received 10 September 2016

Revised 19 September 2017

Accepted 13 October 2017

Available online 2 November 2017

Communicated by Min Xu

Keywords:

Cross-modal retrieval

Multi-label

Multimedia

Partial least squares

ABSTRACT

This paper proposes a novel method named Multi-label Double-layer Learning (MDLL) for multi-label cross-modal retrieval task. MDLL includes two stages (layers): L2C (Label to Common) and C2L (Common to Label). In the L2C stage, considering that labels can provide semantic information, we take label information as an auxiliary modality and apply a covariance matrix to represent label similarity in multi-label situation. Thus we can maximize the correlation of different modalities and reduce their semantic gap in the L2C stage. In addition, we find that samples with the same semantic labels may have different contents from users' view. According to this problem, in the C2L stage, labels are projected to a latent space learned from features of image and text. By this way, the label latent space are more related to the sample's contents. Then, it is noticed that the samples have same labels but various contents can be decreased. In MDLL, iterative learning of the L2C and C2L stages will improve the discriminative ability greatly and decline the discrepancy between the labels and the contents. To show the effectiveness of MDLL, some experiments are conducted on three multi-label cross-modal retrieval tasks (Pascal Voc 2007, Nus-wide, and LabelMe), on which competitive results are obtained.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With the high-speed development of Internet technology, multimedia data has increased dramatically. Consequently, more and more researchers pay their attention on the task of cross-modal retrieval [1–10]. The goal of cross-modal retrieval is to match the feature of one modality with the feature of the other modality in a learned semantic space [11]. To describe the cross-modal retrieval conveniently, we can use an image-text cross-modal retrieval task as an example. In the image-text cross-modal retrieval, given an image query, the text which can describe the image query should be returned; or given a text query, the most related image should be found. The challenging task is that the features of text and image can not be matched directly with each other because implication and numbers of the feature dimensions are both various from modality to modality.

To solve the challenge coming from the heterogeneous feature spaces, one popular solution is to learn a common subspace [2,11–18]. It tries to project the heterogeneous features into the common subspace so as to match the image and text features directly.

Correlations of the image-text pairs are also preserved in subspace learning process. As one of the subspace learning method, Canonical Correlation Analysis (CCA) projects two features of the different modalities to a shared latent space which maximizes the correlations between them [11,19–21]. Besides, Partial Least Squares (PLS) [22–26] is also a classical method of subspace learning, aiming at learning two respective latent spaces by maximizing the correlations between latent spaces.

Cross-modal retrieval can be divided into single-label and multi-label according to the number of labels. Single-label means that each sample belongs to only one semantic class. Previous state-of-art algorithms, such as LGCFL [2], LCFS [5], GMLDA and GMMFA [21], are all based on single-label cross-modal retrieval. However, only one label is not suitable to depict all the objects in the image. Further, in practice, it should allow users to get the retrieval results which are more similar to queries in terms of several semantic classes rather than one. As its particular advantages, the multi-label cross-modal retrieval starts capturing researchers' attention recently. It can describe the samples precisely with the usage of several descriptive labels, and permit users to utilize queries to express their expectation more specifically.

For the multi-label cross-modal retrieval task, this paper proposes a novel approach named Multi-label Double-layer Learning (MDLL). Since labels can preserve the semantic information, MDLL takes label as the auxiliary modality and uses a covariance matrix to present the label similarity in multi-label situation. Then, the

* Corresponding author at : School of Computer and Control Engineering, University of China Academy Science, Beijing, China.

E-mail addresses: jianfeng.he@vipl.ict.ac.cn (J. He), bpma@ucas.ac.cn (B. Ma), shuhui.wang@vipl.ict.ac.cn (S. Wang), liuyg@ucas.ac.cn (Y. Liu), qmh Huang@ucas.ac.cn (Q. Huang).



Label	Image
<i>airport</i> <i>beach</i> <i>sand</i> <i>sun</i> <i>surf</i> <i>water</i>	 

Fig. 1. Illustrative examples show two images in Nus-wide database [27]. Though two images own the same multi-labels, they do not look similar from user's perspective. The image in red box gives the "airport" more weight than other labels, so as the image in blue box gives "beach", "water" and "sand" than others. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

semantic gaps between text and image can be reduced with the introduction of labels. Moreover, with the usage of the semantic information in multi-label, MDLL can improve the performance of cross-modal retrieval greatly. The process using the label information to solve the common space is called as "from Label to Common"(L2C).

In multi-label task, each sample is associated with several labels with the equal weight. Compared with the single-label cross-modal retrieval, the multi-label cross-modal retrieval brings larger within-class similarity. Though some samples own the same labels, they are quite different based on the contents from users' perspective. For instance, second column of Fig. 1 shows two images with the same labels. Their labels are given in the first column of Fig. 1. It can be easily found that two images are not related actually, although they have the same labels. More concretely, the picture in the red box should be set with the larger weight on "airport" than the other five labels. As for the picture in the blue box, it is more relevant to "beach", "sand" and "water" than other three labels. Further, the weight of "beach" is larger than "water", and the weight of "water" is larger than "sand". Thus, we argue that there are some biases in real semantics of the samples because all the labels have the equal weight.

To address the above problem, we consider that labels should be more related to the contents of samples. Firstly, the features, which are extracted from the original multimedia, show more accurate contents of samples compared with the equal-weight labels. Secondly, the contents of samples are reasonable to participate in learning the weights of labels. According to above two points, we learn labels' weight information based on the image's and text's features. Then, with the label latent space substituting the original label space, the influence of above biases can be reduced to some extent. Since each image-text pair should share the same weight for its labels, we use their common space features to update label latent space in our model. The process of solving the label latent space using the common space of samples is called as "from Common to Label"(C2L).

In MDLL, the L2C and C2L stages are conducted in turn. On one hand, in the L2C stage, label information is utilized to solve the common space of the image and text. Through the introduction of label information, the learned common space will preserve the semantic information. On the other hand, the contents of the common space are applied to learn the label latent space. Then the labels have the different weights. With the iterative learning of two stages (layers), the convergence can be found and the goals of two stages can be achieved as possible as they can. In other words, we

achieve that the discriminative ability of the model is improved greatly and the influence of the biases is reduced theoretically.

The novelty and advantages of MDLL can be included as follows:

(1) To solve the multi-label cross-modal retrieval, we propose a novel approach which includes two learning layers: L2C and C2L. By the iterative learning of two layers, the semantic gap is reduced greatly and the discriminative ability improves a lot. To the best of authors knowledge, the iterative learning of double-layer is the first proposed in cross-modal retrieval.

(2) Since multi-label contains richer label information, an extended PLS takes labels as the auxiliary modality and introduces the label information in the form of a covariance matrix of the label latent space. Thus, the relation between the heterogeneous modalities is enhanced in the extended PLS via the covariance matrix.

(3) To reduce the influence coming from the biases between labels and samples' contents, a novel model is designed through enhancing three models' relation in the C2L stage. During this stage, the contents of image and text determines the weight of labels.

The remainder of this paper is organized as follows. In Section 2, we show the extensions of subspace learning method and the state-of-art models in cross-modal retrieval. In Section 3, we give a simple review of PLS. Then, we show the proposed MDLL approach in Section 4. In Section 5, we show the experimental results of MDLL on three public databases. At last, some conclusions are summarized in Section 6.

2. Related work

CCA, as one of the traditional subspace learning methods, has many extensions used in the related area. Based on CCA [11], Semantic Correlation Match (SCM) is proposed to get a semantic subspace by using a logistic regressor. In [4], Correlated Semantic Representation (CSR) obtains a joint image-text representation and an unified formulation by learning a compatible function based on a structural SVM. The 3-view CCA [28], which represents the high-level semantics as a single category or multiple concepts, incorporates the semantics as the third view to solve the cross-modal retrieval problem.

Similar to CCA, PLS has been applied in the cross-modal retrieval problem widely. Sharma et al. [21] and Kang et al. [2] applies PLS to build the relations between the latent variables of image and text. Besides cross-modal retrieval, PLS has been applied successfully on other related problems. For instance, in the task of cross-pose face recognition, the relations between the coupled faces are constructed by PLS [29]. Besides CCA and PLS, Bi-Linear Model(BLM) is also proposed for cross modal face recognition. It is also applied into cross-modal retrieval in [21].

Similarly, there are also many extensions of PLS. In [30], the bridge PLS(BPLS) is proposed by adding ridge-parameter to improve the efficiency of each iteration. Rosipal and Trejo [25] propose kernel PLS(KPLS) by mapping the input variables into a high dimension space so as to solve the nonlinear problem in a linear algorithm. Structured PLS not only learns a low-dimensional and discriminative feature subspace, but also effectively exploits inherently the structural information of labeled image by training data with the structured label information which contained tracking and segmentation simultaneously [31].

In the cross-modal retrieval problem, semantic gaps always exist in the heterogeneous modal spaces. By using label information, semantic gaps can be decreased theoretically [11,32,33]. Specially, GMLDA and GMMFA, constructed to extract multi-view features by a framework based on Generalized Multiview Analysis (GMA), shows the competitive performance on the cross-modal retrieval problem [21]. In [2], label information is used to close the

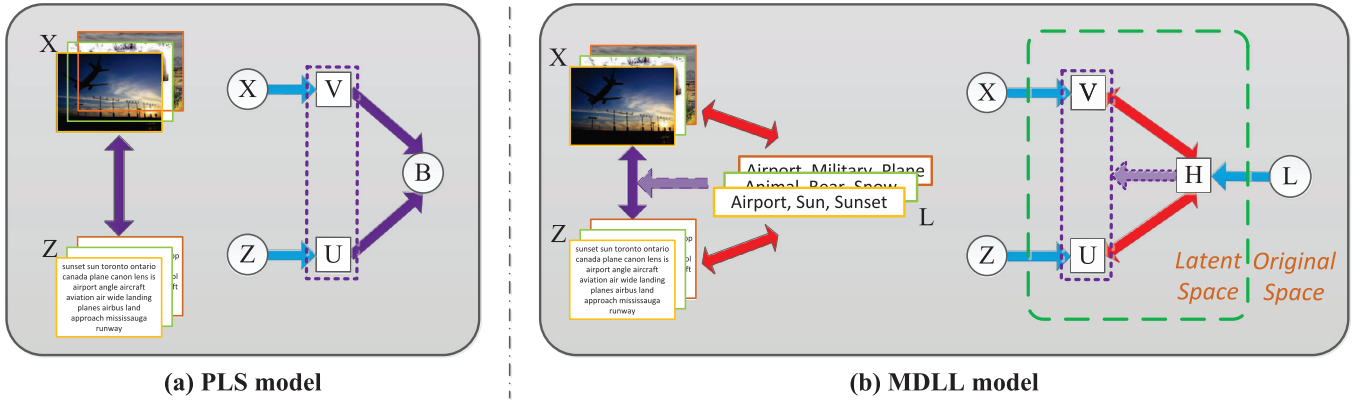


Fig. 2. (a) The structure diagram of PLS. (b) The structure diagram of MDLL. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

different modalities within the same class and enlarge the distances between the heterogeneous modalities. It gains the state-of-art performance on the cross-modal retrieval problem. In [5], Learning Coupled Feature Space(LCFS) is proposed, in which ℓ_{21} -norm is used to select the relevant and discriminative features from the coupled modalities, and trace the norm regularization to enforce the relevances of projected data with potentially connections. It is mentionable that the above proposed methods all aim at the single-label cross-modal retrieval. As for the multi-label cross modal retrieval, Ranjan et al. [34] propose ml-CCA, which is the state-of-art algorithm and utilizes the semantic information in the form of multi-label information and establishes the correspondences across the modalities.

Besides, the databases which can be applied in the multi-label cross-modal retrieval have been built. The popular and applicable databases suitable for the multi-label cross-modal retrieval include PASCAL VOC2007 [35], NUS-WIDE [27] and LabelMe [36].

3. Preliminary

3.1. Multi-label cross-modal retrieval

In this section, we introduce the multi-label cross-modal retrieval. The samples applied in multi-label cross-modal retrieval have single or several labels rather than only one label. The multi-label cross-modal retrieval conforms to actual users' requirement more compared with the single-label. In the multi-label cross-modal retrieval, similar to the single-label cross-modal retrieval, the most related results are returned from one modality with a query from the other modality. To the best of our knowledge, previous researchers of the cross-modal retrieval only focus on the single-label cross-modal retrieval except [34], which for the first time proposes the multi-label cross-modal retrieval.

In this paper, we use $L = [l_1, \dots, l_n]^T \in \mathbb{R}^{n \times c}$ to denote the multi-label indicator matrix, where all the elements in l_i are zeros except for one or several respective semantic classes.

3.2. Partial least squares

PLS can construct the relations between the heterogenous modalities by maximizing the correlation between the latent variables. It has achieved the great successes in many areas [22,23,26]. In Fig. 2(a), we show the structure diagram of PLS.

Let $X = [x_1, \dots, x_n]^T$ represent one multimedia modal original features with n training samples, where x_i is in the space \mathbb{R}^{d_1} . Latent variable of X is represented by $V = [v_1, \dots, v_n]^T \in \mathbb{R}^{n \times p}$ where p is far smaller than d_1 . Similarly, let $Z = [z_1, \dots, z_n]^T \in \mathbb{R}^{n \times d_2}$

represent the original features of the other multimedia modality in the training set. Its latent variable is represented by $U = [u_1, \dots, u_n]^T \in \mathbb{R}^{n \times p}$, where p is also far smaller than d_2 . Finally, PLS can be built as:

$$\begin{cases} X = VW^T + \varepsilon_x \\ Z = UQ^T + \varepsilon_z \end{cases} \quad (1)$$

where the matrices W and Q are the loading matrices, the matrices ε_x and ε_z are the residuals matrices. By means of the low dimension latent variables V and U , we can further get a regression coefficient matrix $B \in \mathbb{R}^{d_1 \times d_2}$ and then project X into Z through B as follow:

$$\begin{cases} B = X^T U (V^T X X^T U)^{-1} V^T Z \\ Z = X B^T + \varepsilon_B \end{cases} \quad (2)$$

where ε_B is the residual matrix. One thing should be pointed out is that the sample data X and Z are Z-score normalized. Thus, their sample covariance matrix $cov(X, Z)$ is as follow:

$$cov(X, Z) = \frac{X^T Z}{n-1} \quad (3)$$

According to Abdi [37], PLS can be solved by a traditional iterative algorithm calculating the first dominant eigenvector to get the weight vectors r and s as follow:

$$\begin{cases} X^T Z Z^T X r = \lambda_1 r \\ Z^T X X^T Z s = \lambda_2 s \end{cases} \quad (4)$$

where λ_1 and λ_2 are respective eigenvalues.

Applying Eq. (3) in the form of $X^T Z = (n-1)cov(X, Z)$ into Eq. (4), it can be found that $Z^T X X^T Z s = [(n-1)cov(X, Z)]^T [(n-1)cov(X, Z)] s = \lambda_2 s$, which means the weight vectors r and s are corresponding to the first right singular vector and the first left singular vector of $(n-1)cov(X, Z)$. However, in the SVD, the coefficient of a matrix only affects the middle term which is a diagonal matrix and has no effect on the left singular vectors and right singular vectors. Thus, the weight vectors r and s can be derived from the SVD of $cov(X, Z)$ as follow:

$$cov(X, Z)^T cov(X, Z) s = \lambda_2 s \quad (5)$$

After the i th iteration, we can obtain the i th latent vectors $v_i = X r_i$ and $u_i = Z s_i$. Then we get the weight matrices R and S , followed by solving the latent variables V and U . Finally, the relation between X and Z is built by Eq. (2) which means X and Z are comparable.

4. The proposed MDLL model

In this section, MDLL and its optimization are introduced in details. The structure diagram of MDLL is shown in Fig. 2 (b), in

which the purple arrows are the L2C stage, the red arrows are the C2L stage, and the blue arrows are the projection.

4.1. From label to common (L2C)

Compared with the heterogeneous features of the different modalities, label information is more related to the semantic information. So, for the semantic gap between text and image in the cross-model retrieval, we argue that label information are effective to build the relation between text and image. Thus, similar to [28,38], we regard labels as the auxiliary modality to reduce the gap between the heterogeneous modalities.

In the L2C stage, we learn a more discriminate common space by designing a sample covariance matrix ψ_{xz} which contains label similarity as follows:

$$\psi_{xz} = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^n \xi(l_i, l_j) x_i z_j^T \quad (6)$$

where $N = n \times n$ and ξ describes the similarity between the label vectors l_i and l_j :

$$\xi(l_i, l_j) = \exp(-\|l_i - l_j\|_2^2 / \sigma) \quad (7)$$

where σ is a constant factor.

According to Eq. (5), in the L2C stage, the sample covariance matrix $cov(X, Z)$ which is utilized in SVD to solve PLS is replaced by ψ_{xz} . Hence, similar to Eq. (5), the object function of L2C stage to solve weight vector s is as follow:

$$\begin{aligned} \psi_{xz}^T \psi_{xz} s &= \lambda_3 s \\ \text{s.t. } s^T s &= 1 \end{aligned} \quad (8)$$

where λ_3 is eigenvalue.

From above object function, we can conclude that during the process solving modal common space by PLS, the L2C stage not only achieves maximizing the correlation between the latent variables $v(v = Xr)$ and $u(u = Zs)$, but also introduces the label information. On the one side, maximizing the correlation between the latent variables enables the features of the different modalities more comparable after projection. On the other side, introducing the label information can reduce the semantic gaps theoretically. More meaningfully, the usage of the covariance matrix ψ_{xz} has avoided the question which label should be selected in using the label information.

To solve Eq. (8), we use the SVD described in Section 3 by solving the weight vectors r and s which are respective the first right singular vector and the first left singular vector of ψ_{xz} . Because we can solve both r_i and s_i at the i -th iteration, we just give the object function solving weight vector s without the other one about r .

4.2. From common to label (C2L)

As introduced in Section 1, the equal weight of the multi-labels may lead to some biases in the retrieval results from users' view. In allusion to this problem, we argue that the labels should have the weight divergence in terms of the contents. Thus, in the C2L stage, the original labels are projected into a latent space. We learn the label latent space with the help of the features of the image's and text's common space. They are learned from their original features which include more accurate contents compared with the equal weight labels. With the extra information about the contents, the label latent space is more accurate to describe the sample contents. In other words, the weights for each label are no more equal in the label latent space. By this way, the influence of the equal weight can be remitted to certain extent.

In the C2L stage, the operator $P = [p_1, \dots, p_c]^T \in \mathbb{R}^{c \times p}$ projects the labels into the label latent space. Thus, by closing the distance between the label latent space and the feature latent spaces,

we can enhance the relation between the features of the different modalities and their labels. Furthermore, based on the Frobenius-norm, the projection operator P can be learned as:

$$\arg \max_P \beta = -\frac{1}{2} (\|XR - LP\|_F^2 + \|ZS - LP\|_F^2 + \mu \|P\|_F^2) \quad (9)$$

where μ is a nonnegative regularization parameter. In the first two term of Eq. (9), P is constrained by the common space of the image and text to achieve that the weights of labels are more related to the sample contents and no more equal. The last term $\mu \|P\|_F^2$ is the transformation constraint to prevent overfitting.

To solve Eq. (9), we fix R and S to update P . After setting $\frac{d\beta}{dR}$ to 0, we can obtain:

$$\begin{aligned} (2L^T L + \mu I)P &= L^T X R + L^T Z S \\ \Rightarrow P &= (2L^T L + \mu I)^{-1} (L^T X R + L^T Z S) \end{aligned} \quad (10)$$

4.3. Solve MDLL by iteration

In the C2L stage, the label latent space is updated under the constraint of the common space. Thus the label information applied in model is more relative to the contents of samples. However, the change of the label latent space in each iteration leads to various input of Eq. (8). In the L2C stage, the updated label latent space makes the common space more discriminated for its extra content information. Because of the various input of Eq. (8), the L2C stage then outputs different modal common space in each iteration. With the change of the common space, the input of the C2L is altering. And above is what has changed in each iteration.

The concrete procedures of MDLL can be found in Alg. 1. We achieve the iterative learning of two modalities and labels with two objective functions Eq. (8) and Eq. (9). Specially, in the L2C stage, the labels L can be updated by

$$L' = LP \quad (11)$$

followed by Eq. (7) updated and then the R and S is also updated. In the C2L stage, we solved the label projection operator P via Eq. (10).

Finally, we achieve the semantic gap declined conspicuously and the influence caused by the equal-weight of the labels dropped to certain extent with the iterative learning of the modal common space and the label latent space.

4.4. Analysis of computational complexity

Lastly, we briefly analyze the computational complexity of MDLL.

In the L2C stage, MDLL solves the modal latent space. The latent variables are p dimensions. And each dimension is solved by the SVD calculating the first right singular vector and the first left singular vector. Set n as the number of sample pairs in the training set, as a result of d_1 dimensions for image features and d_2 dimensions for it is text features. And we assume $d_1 > d_2$, then the computational complexity of the L2C stage is $O(pd_1 d_2^2)$.

In the C2L stage, MDLL solves label projection operator. In this part, the time complexity of the first term in Eq. (10) is $O(c^3)$, and the time complexity of the second term in Eq. (10) is $O(cnp(d_1 + d_2))$. Thus the time complexity of the C2L stage is $O(c^3 + cnp(d_1 + d_2))$.

Finally, as a result of combining the L2C and C2L stages in iterations where the max iteration is set as m , the time complexity of MDLL is $O(m(pd_1 d_2^2 + (c^3 + cnp(d_1 + d_2))))$.

5. Experiments

In this section, we test MDLL on three popular databases to show its performance on multi-label cross-modal retrieval task.

Algorithm 1 The algorithm of MDLL approach.**Input:**

the different feature modality X and Z , the class indicator matrix L , the dimension p of latent space, the coefficient μ of regularization term

Output:

Converged P, R, S

- 1: Initialize P using identify matrix;
- 2: Store X and Y as $E = X, F = Z$;
- 3: **repeat**
- 4: Update L with Eq. (11);
- 5: Update ψ_{xz} with Eq. (6);
- 6: **for** $i = 1$ to p **do**
- 7: Calculate the first right singular vector and the first left singular vector of ψ_{xz} to obtain the i th weight vector r_i and s_i correspondingly;
- 8: Calculate the i th latent vector:
 $v_i = Xr_i, u_i = Zs_i$;
- 9: Deflate X, Z matrices as follow:
 $X = X - v_i v_i^T X, Z = Z - u_i u_i^T Z$;
- 10: **end for**
- 11: Update $X = E$ and $Z = F$
- 12: Calculate the P with Eq. (10)
- 13: **until** Convergence criterion satisfied
- 14: **return** P, R and S ;

5.1. Experimental databases

Nus-wide [27] is crawled from Flickr website including 269,648 image-tag pairs. There are 81 semantic concepts for this database, which can be regarded as the class labels in our experiment. Each image is annotated with one or several labels of 81 semantic concepts. The class labels which correspond top-10 largest numbers of image are picked out for our experiment. As a result out of the top-10 largest numbers of image, we choose 67,993 image-tag pairs. Then, we randomly select 40,834 image-text pairs for the training set and 27,159 image-text pairs for the testing set. For feature representations, we use the 500-dimensional bag-of-words vectors based on the SIFT descriptors as the image features and the 1000-dimensional word frequency vectors based on tag features as the text features.

Pascal VOC 2007 [35] consists of 5011/4952 (training/testing) image-tag pairs with 20 semantic classes. For the feature representation, we use the 512-dimensional GIST features for image, and the 399-dimensional absolute tag rank features for text. For the label representation, we use its semantic classes. We use the original train-test split provided in the database for training and testing.

LabelMe includes 3825 image collected by Hwang and Grauman [36]. We use the publicly available GIST features provided by Hwang and Grauman [36] for image representation. For it is text representation, we use the 209 dimensional absolute tag rank features provided by Hwang and Grauman [36]. For label representation, we use the groundtruth annotation of the image. We perform a random 50 to 50 split of the database for creating the training and testing sets.

5.2. Compared scheme

To validate the performance of MDLL in multi-label cross-modal retrieval, we compare it with one baseline and several related state-of-art approaches. CCA is a traditional subspace learning approach which projects heterogeneous modal features into a shared latent space by maximizing the correlation between two modalities. We use CCA as a baseline algorithm. LCFS unifies linear pro-

Table 1
MAP results on the VOC database.

Methods/Tasks	txt2im	im2txt	Average
CCA [26]	0.3073	0.2945	0.3009
LCFS [5]	0.4278	0.3355	0.3816
LGCFI [2]	0.4362	0.3440	0.3901
ml-CCA [34]	0.4280	0.3584	0.3932
MDLL	0.4604	0.3745	0.4174

Table 2
MAP results on the NUS-WIDE database.

Methods/Tasks	txt2im	im2txt	Average
CCA [26]	0.2869	0.2667	0.2768
LCFS [5]	0.4742	0.3363	0.4053
LGCFI [2]	0.4972	0.3907	0.4440
ml-CCA [34]	0.4689	0.3908	0.4299
MDLL	0.4874	0.4037	0.4455

jection operators, ℓ_{21} norm and trace norm to learn a subspace and select coupled features simultaneously. LGCFI is a supervised method regarding the multimedia modalities as assemblies of local parts to learn the most discriminant groups. ml-CCA uses the semantic information to establish the correspondences in the form of multi-label information.

5.3. Evaluation metric

In our experiment, we use MAP (Mean Average Precision) and PR (Precision-Recall) curve to show the effectiveness of MDLL.

MAP has been widely used to evaluate the overall performance of cross-modal retrieval, such as [2,6,11,34,39]. To compute MAP, we first evaluate the average precision (AP) of a retrieved database including N retrieved samples by $AP = \frac{1}{T} \sum_{r=1}^N E(r) \delta(r)$, where T is the number of the relevant samples in the retrieved database, $E(r)$ denotes the precision of the top r retrieved samples, and $\delta(r)$ is set to 1 if the r th retrieved sample is relevant (on above three databases, a retrieved sample is relevant if it shares at least one label with the query) and $\delta(r)$ is 0 otherwise. Then by averaging the AP values over all the queries, MAP can be calculated.

Besides, PR curve is a classical measure of information retrieval or classified performance. Assume that the set S_1 includes the samples in which real labels are denoted by L_r . The classifier picks out the set S_2 samples in which labels are classified into L_r . In the set S_2 , the samples in which real labels are L_r construct the set S_3 . Thus, we can calculate the precision ratio: $PR = \frac{|S_3|}{|S_2|}$ and the recall ratio: $RR = \frac{|S_3|}{|S_1|}$, where $|A|$ means the number of elements in set A . Furthermore, we get different PR-RR values via the different classified setting and then draw precision-recall curve in which the vertical coordinate is precision ratio and the horizontal coordinate is recall ratio.

5.4. Experimental results

The MAP values of all the algorithms on the Pascal Voc2007, NUS-WIDE and LableMe databases are presented in Tables 1, 2 and 3, respectively. The results significantly better than others are indicated in boldface.

5.4.1. Results on Pascal VOC 2007

From Table 1, we have the following observations: firstly, MDLL achieves better performances than other compared algorithms in terms of MAP. It is mentionable that the other algorithms all just

Table 3
MAP results on the LabelMe database.

Methods/Tasks	txt2im	im2txt	Average
CCA [26]	0.5656	0.5753	0.5704
LCFS [5]	0.7898	0.8067	0.7982
LGCFL [2]	0.8390	0.7961	0.8176
ml-CCA [34]	0.8175	0.8081	0.8128
MDLL	0.8601	0.8559	0.8580

use the single stage but MDLL owns two stages. It indicates the dominance of the iterative learning of the L2C and C2L stages.

Secondly, LCSF, LGCFL, ml-CCA and MDLL outperform at least 27.12% higher average MAP than CCA. Considering that these methods take label information into the model while CCA only use the multimedia modal features, it is believable to conclude that the label information can provide valuable information in multi-label cross-modal retrieval.

Third, MDLL achieves 9.38% higher average MAP than LCSF. LCSF uses the original label space to solve the multimedia modal common space. Different from LCSF, MDLL makes use of label in computation of the common space. It illustrates effectiveness of the introduction of the labels' latent space.

Moreover, MDLL outperforms 7% higher average MAP than LGCFL. Similarly, LGCFL also uses the original label space to solve the projection operators. It indicates the benefit of labels' latent space in our approach again. In addition, LGCFL does not take the similarity between labels into account while MDLL does. Thus, it is valid to consider the similarity between labels in the L2C stage.

At last, compared with ml-CCA which also uses the similarity between the corresponding multi-label vectors to show the relation between the heterogeneous modal data, MDLL obtains 6.15% higher average MAP. It demonstrates the necessity and advantage of using the latent space of the labels. By learning the labels' latent space combined with multimedia modal features, we enhance the relation between three modalities in the C2L stage. Thus they can avoid the over fitting because the variation from practical contents to tagged label, which achieves more discriminative ability in matching the semantic-similar heterogeneous pairs.

In Fig. 3, MDLL also has the best performance on both image-to-text and text-to-image tasks in Pascal Voc 2007. From the PR curves, it is obvious that under the same recall rate, MDLL gets

the highest precision in all compared algorithms. That also shows superiority of MDLL.

5.4.2. Results on NUS-WIDE

According to the Table 2, we get the conclusion as follow:

To begin with, MDLL also outperforms the compared state-of-the-art algorithms in terms of the average MAP. Compared with the second best algorithm LGCFL, MDLL gets the average MAP 1.13% increased. Though the improvement is only 1.13%, it still shows the competitive result of MDLL on multi-label cross-modal retrieval.

Besides, we can also find that the MAP of image-to-text of MDLL is the highest result among the five algorithms. This may be the effect of the C2L which learns more accurate contents from the model features rather than the conventional labels. Thus, the common space is more consistent compared with others learned by other algorithms.

Nevertheless, different from the results based on the Pascal VOC database, LGCFL outperforms 2.1% higher text-to-image MAP than MDLL. The possible reason for that result is that the image in LGCFL are employed as regularization in the form of assemblies of local parts set, which means the image information is applied in a more effective way.

In Fig. 4, it is clear that MDLL achieves the best performance on the text-to-image query task. As for the image-to-text query task, MDLL ranks the second in some case.

5.4.3. Results on LabelMe

In Table 3, MDLL reaches the best overall performance in average MAP again, from which, we can conclude as follow:

Firstly, it exceeds existing state-of-art algorithms with MAP 81.76% for LGCFL and 81.28% for ml-CCA. The improvement of MDLL in average MAP is at least 4.94% on this database. It validates the effectiveness of MDLL one more time.

Secondly, the performance of supervised algorithms is still far higher than unsupervised algorithm(CCA). This shows the importance of labels in cross-modal retrieval again.

Thirdly, it is obvious that the results on this database are generally high. We attribute these supernal figures to its less classes. To be more specific, the LabelMe collected by Hwang and Grauman [36] includes only 5 categories(person, car, screen, keyboard, and mug). This leads to more distinguish labels than other databases. Then, the retrieval difficulty plunges.

In Fig. 5, the PR curves are also displayed. It shows that MDLL ranks top on the text-to-image task. As for image-to-text task, it is

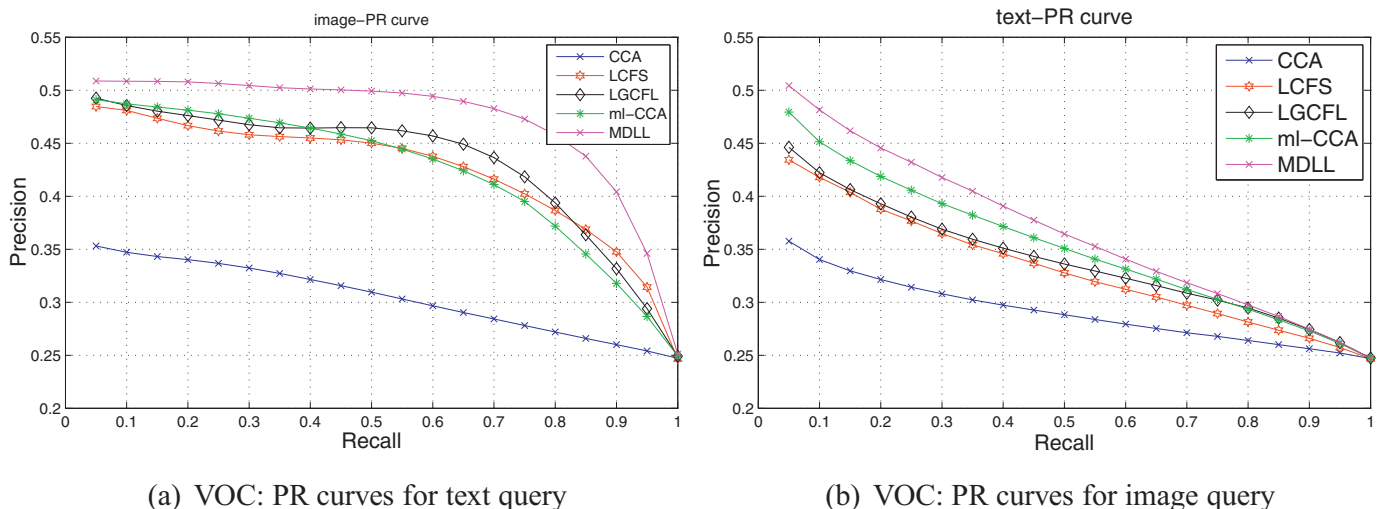


Fig. 3. PR curves on the VOC database.

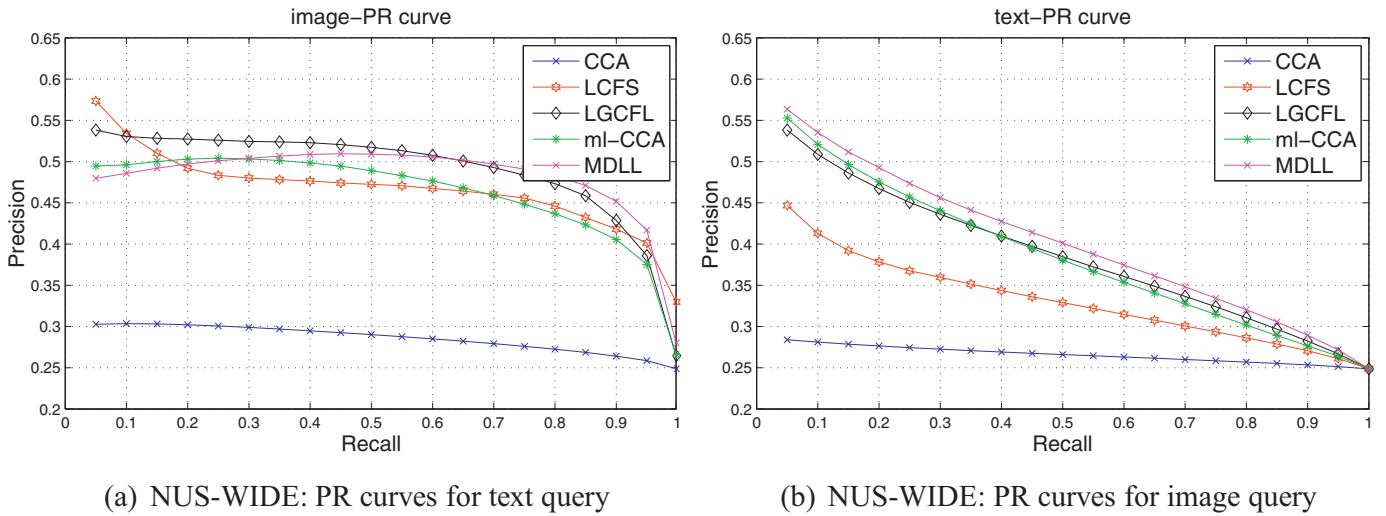


Fig. 4. PR curves on the NUS database.

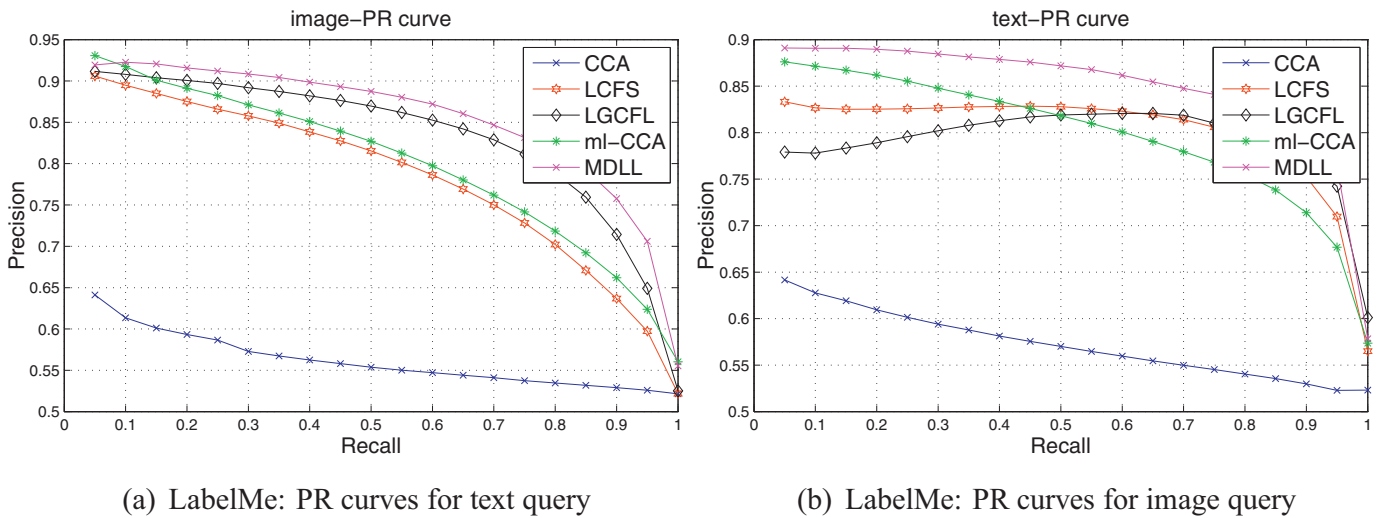


Fig. 5. PR curves on the LabelMe database.

mentionable that the PR curve of our method is much better than any other methods at middle and high levels of recall, which is more practicable in high-performance required retrieval.

5.5. Parameter sensitivity

In MDLL, we need to tune several parameters. In our experiments, we set constant factor σ to 100 for three databases. To validate how the rest parameters affect the performance, we repeat the experiments to evaluate the parameter sensitivity on the PASCAL VOC2007 database.

Fig. 6 shows the MAP values in the process tuning parameters p . p is the number of the latent space dimension of multimedia modalities. From that, we find that MAP is stable when p is bigger than 100, which indicates that the dimension number of latent space should be large enough. After considering the time complexity of algorithm, we set p to 100 for PASCAL VOC 2007. Besides, we also set $p = 200$ and $p = 100$ for Nus-wide and Labelme respectively. Furthermore, we set regularization parameter $\mu = 0.1$ for Pascal Voc 2007, $\mu = 1$ for both Nus-wide and LabelMe databases and max iteration $m = 20$ for the all databases.

Table 4

Calculational time on the VOC database.

Methods	Tasks	
	Training time	Testing time
CCA	52.82	10.69
LCFS	166.53	10.76
LGCFL	36.62	10.52
ml-CCA	245	10.82
MDLL	2339.10	10.87

5.6. Comparison on computation time

In addition, we also carry on the comparison on computation time. All algorithms run on the Pascal Voc 2007 database based on the setting proposed in Section 5.1 and Section 5.5. Moreover, the computational time includes training time and testing time as Table 4, in which the unit is second.

It is noticed that our model is much more time-consuming than the state-of-the-art algorithms in the terms of training time. Because our model has two layers and each layer calculates

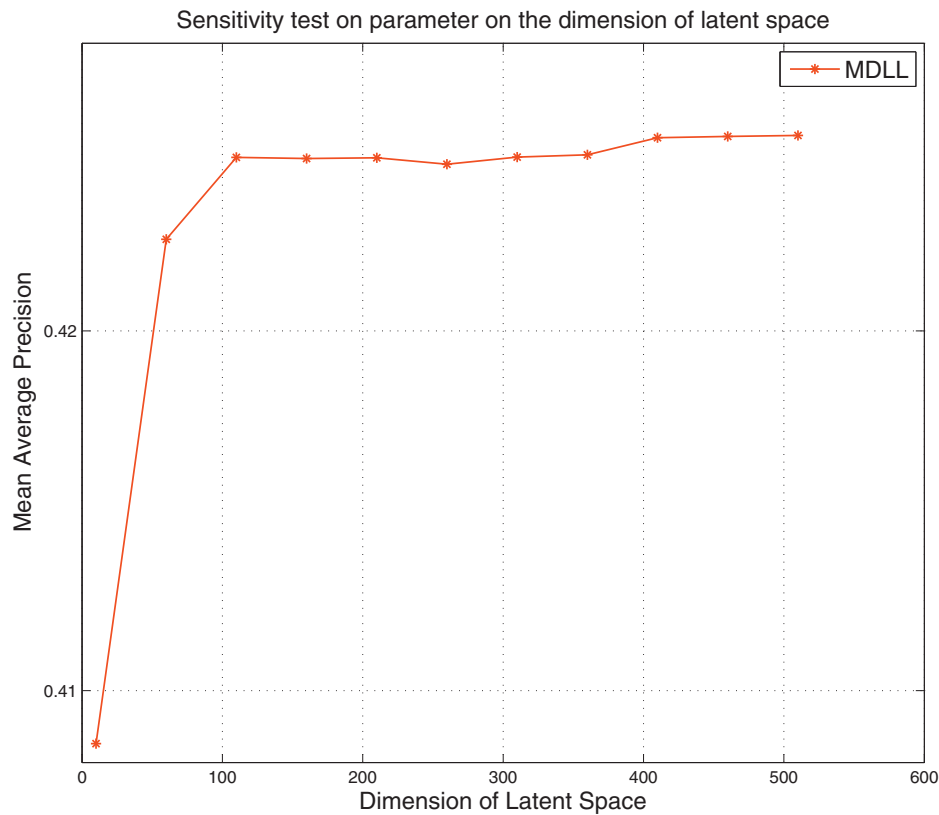


Fig. 6. Sensitivity analysis of dimension of the latent space on PASCAL VOC2007.

Text Query	Image Satisfied Multi-Tag Query					Method	Image Query	Image Satisfied Multi-Tag Query					Method
airplane sky						CCA							CCA
						LCFS							LCFS
						ml-CCA							ml-CCA
						LGCFI							LGCFI
						MDLL							MDLL
bird fence water						CCA							CCA
						LCFS							LCFS
						ml-CCA							ml-CCA
						LGCFI							LGCFI
						MDLL							MDLL

Fig. 7. Two examples of queries and their results retrieved by MDLL on the Pascal VOC2007 database.

iteratively. The enormous iteration leads to the obviously higher training time. As for the testing time, it is obvious that each algorithm has similar testing time. Since they all project the original modal features into common space by respective projection operators in the test processes. Nevertheless, the training is done offline and only once. Thus the training time cost is not as important as the testing time. In the future, we will improve the efficiency of the proposed work.

5.7. Exhibition of retrieval result

Besides above experiments, we also show instances of queries and their results retrieved by MDLL and other comparing algorithms on the Pascal Voc 2007 database in Fig. 7. In the left subfigure, a text query and its respective images of ground truth are

shown in the first column. The top five retrieved images of MDLL and other comparing algorithms are exhibited from the second column to the sixth column. Red frames are wrong retrieval results based on their respective labels. From the left subfigure, we can know that all the retrieved images of MDLL are correct while at least one result is wrong in all other algorithms. In the right subfigure, similarly, we also show the image query retrieval results including its respective text of ground truth, the top five retrieved documents represented by their corresponding image for all algorithms. From that, we also can find that at least one retrieval result is wrong for other algorithms.

To sum up, through the above MAP tables, PR curves, and exhibition of retrieval results, MDLL gets competitive results on multi-label cross-modal retrieval because of the iterative learning of the L2C and C2L stages.

6. Conclusion

This paper proposes a novel method for the multi-label cross-modal retrieval task. In our approach, we utilize iterative update of two stages: L2C and C2L. The L2C stage maximizes the correlation of two multimedia modal latent variables with the information of label in the common space, and the C2L stage reinforces the relation between three modalities through learning labels' latent space based on modal features showing samples' contents. The experiments are carried out on three public databases, validating that MDLL outperforms the state-of-the-art methods.

Later, we will look for more practical methods to evaluate performance of multi-label cross-modal retrieval and apply deep neural network into multi-label cross-modal retrieval.

Acknowledgments

This work was supported in part by National Basic Research Program of China (973 Program): 2015CB351800, in part by Natural Science Foundation of China (NSFC): 61332016, 61572465, 61672497, 61620106009, U1636214 and 61650202, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

References

- [1] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (3) (2014) 521–535.
- [2] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, *IEEE Trans. Multimed.* 17 (3) (2015) 370–381.
- [3] X. Mao, B. Lin, D. Cai, X. He, J. Pei, Parallel field alignment for cross media retrieval, in: *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 897–906.
- [4] Y. Verma, C. Jawahar, Im2text and text2im: associating images and texts for cross-modal retrieval, in: *Proceedings of the British Machine Vision Conference*, 2014.
- [5] K. Wang, R. He, W. Wang, L. Wang, T. Tan, Learning coupled feature spaces for cross-modal matching, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2088–2095.
- [6] L. Xie, P. Pan, Y. Lu, A semantic model for cross-modal and multi-modal retrieval, in: *Proceedings of the ACM Conference on International Conference on Multimedia Retrieval*, 2013, pp. 175–182.
- [7] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, W. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 2013.
- [8] T. Yao, X. Kong, H. Fu, Q. Tian, Semantic consistency hashing for cross-modal retrieval, *Neurocomputing* 193 (2016) 250–259.
- [9] S. Wang, F. Zhuang, S. Jiang, Q. Huang, Q. Tian, Cluster-sensitive structured correlation analysis for web cross-modal retrieval, *Neurocomputing* 168 (2015) 747–760.
- [10] H. Zhang, Y. Liu, Z. Ma, Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval, *Neurocomputing* 119 (2013) 10–16.
- [11] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: *Proceedings of the ACM International Conference on Multimedia*, 2010, pp. 251–260.
- [12] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinomial data, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2407–2414.
- [13] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, *Neural Comput.* 12 (6) (2000) 1247–1283.
- [14] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Cross-modal retrieval using multi-ordered discriminative structured subspace learning, *IEEE Trans. Multimed.* 19 (6) (2017) 1220–1233.
- [15] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, PI-ranking: a novel ranking method for cross-modal retrieval, in: *Proceedings of the ACM International Conference on Multimedia*, 2016, pp. 1355–1364.
- [16] L. Zhang, B. Ma, J. He, G. Li, Q. Huang, Q. Tian, Adaptively unified semi-supervised learning for cross-modal retrieval, in: *Proceedings of the International Conference on Artificial Intelligence*, 2017, pp. 3406–3412.
- [17] X. Bai, S. Bai, Z. Zhu, L. Latecki, 3d shape matching via two layer coding, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (12) (2015) 2361–2373.
- [18] S. Bai, X. Bai, Sparse contextual activation for efficient visual re-ranking, *IEEE Trans. Image Process.* 25 (3) (2016) 1056–1069.
- [19] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods, *Neural Comput.* 16 (12) (2004) 2639–2664.
- [20] A. Li, S. Shan, X. Chen, W. Gao, Maximizing intra-individual correlations for face recognition across pose differences, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 605–611.
- [21] A. Sharma, A. Kumar, H.D. III, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2160–2167.
- [22] B. Ding, R. Gentleman, Classification using generalized partial least squares, *J. Comput. Graph. Stat.* (2012) 280–298.
- [23] M.A. Haj, J. Gonzalez, L. Davis, On partial least squares in head pose estimation: how to simultaneously deal with misalignment, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2602–2609.
- [24] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, *Lecture Notes in Computer Science* 3940 (2006) 34–51.
- [25] R. Rosipal, L.J. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert space, *J. Mach. Learn. Res.* 2 (2002) 97–123.
- [26] A. Sharma, D.W. Jacobs, Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 593–600.
- [27] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, p. 48.
- [28] Y. Gong, Q. Ke, M. Isard, S. Lazebnik, A multi-view embedding space for modeling internet images, tags, and their semantics, *Int. J. Comput. Vis.* 106 (2) (2014) 210–233.
- [29] A. Li, S. Shan, X. Chen, W. Gao, Cross-pose face recognition based on partial least squares, *Pattern Recognit. Lett.* 32 (15) (2011) 1948–1955.
- [30] J. Tang, H. Wang, Y. Yan, Learning hough regression models via bridge partial least squares for object detection, *Neurocomputing* 152 (2015) 236–249.
- [31] B. Zhong, X. Yuan, R. Ji, Y. Yan, Z. Cui, X. Hong, Y. Chen, T. Wang, D. Chen, J. Yu, Structured partial least squares for simultaneous object tracking and segmentation, *Neurocomputing* 133 (2014) 317–327.
- [32] Y. Gong, S. Lazebnik, Iterative quantization: a procrustean approach to learning binary codes, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 817–824.
- [33] J. Wang, S. Kumar, S. Chang, Semi-supervised hashing for large-scale search, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (12) (2012) 2393–2406.
- [34] R. Viresh, R. Nikhil, J. CV, Multi-label cross-modal retrieval, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.
- [35] M. Everingham, L.V. Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge 2007 (VOC 2007) results (2007), 2008. <http://www.pascalnetwork.org/challenges/VOC/voc2008/workshop/index.html>
- [36] S.J. Hwang, K. Grauman, Reading between the lines: object localization using implicit cues from image tags, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (6) (2012) 1145–1158.
- [37] H. Abdi, Partial least squares regression and projection on latent structure regression (PLS regression), *Wiley Interdiscip. Rev. Comput. Stat.* 2 (1) (2010) 97–106.
- [38] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and semisupervised regularization, *IEEE Trans. Circuits Syst. Video Technol.* 24 (6) (2014) 965–978.
- [39] G. Song, S. Wang, Q. Huang, Q. Tian, Similarity Gaussian process latent variable model for multi-modal data analysis, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4050–4058.



Jianfeng He received the B.E. degree in digital media from Central China Normal University, Wuhan, China, in 2014, and received the M.S. degree in computer technology at the University of the Chinese Academy of Sciences, Beijing, China, in 2017. His research interests include machine learning, image content analysis, and information retrieval.



Bingpeng Ma received the B.S. degree in mechanics in 1998 and the M.S. degree in mathematics in 2003 from Huazhong University of Science and Technology, respectively. He received Ph.D. degree in computer science at the Institute of Computing Technology, Chinese Academy of Sciences, P.R. China in 2009. He was a post-doctoral researcher in University of Caen, France, from 2011 to 2012. He joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, in March 2013 and now he is an assistant professor. His research interests cover image analysis, pattern recognition, and computer vision.



Shuhui Wang received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012. He is an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include large-scale Web data mining, visual semantic analysis and machine learning.



Yugui Liu received B.S. degree in mathematics from Peking University, Beijing, China, in 1984, and the M.S. degree in Computer Science from Graduate University of Chinese Academy of Sciences, Beijing, China, in 1996. He is an associate professor with the School of Computer and Control Engineering, Chinese Academy of Sciences. His research interests include multimedia technology and distributed multimedia system.



Qingming Huang is a professor in the University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. He was a Postdoctoral Fellow in the National University of Singapore from 1995 to 1996, and served as a member, research staff in the Institute for Infocomm Research, Singapore from 1996 to 2002. He joined the University of Chinese Academy of Sciences as a professor under the Science 100 Talent Plan in 2003, and has been granted by China National Funds for Distinguished Young Scientists in 2010. He is also the recipient of the National Hundreds and Thousands Talents Project in 2014. His research areas include multimedia video analysis, image processing, computer vision and pattern recognition. He has published more than 300 academic papers in prestigious international journals including IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, IEEE Trans. on CSVT, etc. and top-level conferences such as ACM Multimedia, ICCV, CVPR, IJCAI, VLDB, etc. He is the associate editor of Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on Multimedia, IEEE Trans. on CSVT, IEEE Trans. on Image Processing, etc. He is a senior member of IEEE and has served as program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PSIVT, etc.