Multimed Tools Appl
https://doi.org/10.1007/s11042-018-5830-y

Combinative hypergraph learning in subspace for cross-modal ranking

Fangming Zhong[1] · Zhikui Chen[1] · Geyong Min[2] · Zhaolong Ning[1] · Hua Zhong[1] · Yueming Hu[3]

Received: 3 July 2017 / Revised: 27 December 2017 / Accepted: 21 February 2018

Abstract Recent years have witnessed a surge of interests in cross-modal ranking. To bridge the gap between heterogeneous modalities, many projection based methods have been studied to learn common subspace where the correlation across different modalities can be directly measured. However, these methods generally consider pair-wise relationship merely, while ignoring the high-order relationship. In this paper, a combinative hypergraph learning in subspace for cross-modal ranking (CHLS) is proposed to enhance the performance of cross-modal ranking by capturing high-order relationship. We formulate the cross-modal ranking as a hypergraph learning problem in latent subspace where the high-order relationship among ranking instances can be captured. Furthermore, we propose a combinative hypergraph based on fused similarity information to encode both the intra-similarity in each modality and the inter-similarity across different modalities

✉ Fangming Zhong
fmzhong@mail.dlut.edu.cn

Zhikui Chen
zkchen@dlut.edu.cn

Geyong Min
g.min@exeter.ac.uk

Zhaolong Ning
zhaolongning@dlut.edu.cn

Hua Zhong
zhonghua@mail.dlut.edu.cn

Yueming Hu
ymhu163@163.com

[1] School of Software Technology, Dalian University of Technology, Dalian, China

[2] College of Engineering, Computing and Mathematics, University of Exeter, Exeter, UK

[3] College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China

Published online: 05 March 2018

into the compact subspace representation, which can further enhance the performance of cross-modal ranking. Experiments on three representative cross-modal datasets show the effectiveness of the proposed method for cross-modal ranking. Furthermore, the ranking results achieved by the proposed CHLS can recall 80% of the relevant cross-modal instances at a much earlier stage compared against state-of-the-art methods for both cross-modal ranking tasks, i.e. image query text and text query image.

Keywords Cross-modal ranking · Subspace learning · Hypergraph · Similarity preserving

1 Introduction

Currently, the investigation of multiple views and modalities in computer vision tasks, such as visual tracking [12–14], object recognition [15, 16, 18, 19], and visual retrieval [21, 32], have attracted many research interests, especially the cross-modal retrieval problem [10, 36, 42]. The various modal data existing on the Web, such as image, text and video, provide complementary information to describe the semantics of objects. The ranking of cross-modal retrieval enhances the results of cross-modal retrieval. For instance, given a text query, the top-k close images should be returned with their relevance scores in the descending order or given an image query, the top-k close textual documents should be returned [26]. Hence, cross-modal ranking is essential for cross-modal retrieval. However, the heterogeneity gap among different modalities is still a challenging problem for the cross-modal ranking in distance metric learning. Another main challenge in cross-modal ranking is how to capture the high-order relationship among samples and use the complementary intra-modal similarity to enhance cross-modal ranking.

In recent years, a number of studies [3, 34, 37, 38, 50] have been conducted to bridge the heterogeneity gap between different modalities e.g., text and image. These approaches can be categorized into two main groups: 1) subspace learning either in unsupervised or supervised manner, and 2) cross-modal hashing.

Subspace learning methods aim to learn a latent subspace, in which the similarity among different modalities can be measured directly. The unsupervised subspace learning methods such as Canonical Correlation Analysis (CCA) [26, 41], Partial Least Squares (PLS) [28], and Locality Preserving Projections (LPP) [5] map the multimodal data into a common space in which they are highly correlated. In contrast, the supervised learning methods [25, 54] utilize the class label information to obtain more discriminative subspace. For example, if two samples have the same class label, their projections should be as close as possible. Otherwise, if they have different class labels, their projections should be as far as possible from each other. As the supervised approaches require a number of labeled samples which are fairly expensive to obtain, the semi-supervised subspace learning using both labeled and unlabeled data has attracted increasing attention [4, 49].

Cross-modal hashing (CMH) combines the cross-modal analysis and hashing technology [11, 45, 46, 53]. CMH obtains a unified hash space by learning a set of hashing functions. In the hash space, Hamming distance of hash codes is used to measure the similarity between two different modalities. The strength of CMH methods is that they can use a compact code for multimodal data representation which leads to a low computational complexity. However, the weakness shared by CMH and subspace learning methods is that they only consider the pair-wise relationship between two samples, ignoring the high-order relationship among more than two samples. Generally, we use pair-wise relationship between two items to measure the relationship of them. However, the high-order relationship presents the relationship

among more than two items which are connected. For example, pair-wise relationship cannot define the similarity among three item, it can just describe the closeness of every two items. However, the high-order relationship defines the closeness of three or more items, which can facilitate cross-modal correlation learning [40, 43, 48].

To address this issue, hypergraph was proposed to capture the high-order relationship of samples, which has been widely used in clustering, classification, and information retrieval tasks [43, 48, 52]. In a hypergraph, an edge connects more than two vertices, thus it can well encode the relationship among more than two vertices. Existing studies [40, 48] have shown that the hypergraph is beneficial for multimodal relationship encoding. In this paper, we formulate the cross-modal ranking as a hypergraph learning problem in a common subspace. In order to bridge the semantic gap between different modalities, the original modalities are projected onto a common subspace, where the distance among heterogeneous modalities can be measured. Here, Canonical Correlation Analysis (CCA) is utilized for latent subspace learning. CCA can well preserve the correlations in paired samples, owning to the computation of maximum correlation coefficient. Then we employ the hypergraph learning to compute cross-modal ranking scores in the common subspace. To further improve the cross-modal ranking performance, a combinative hypergraph is constructed which takes into consideration both the intra-similarity in each modality and the inter-similarity across different modalities. Therefore, by performing hypergraph learning in the subspace, our approach captures not only the pair-wise but also the high-order relationship among ranking objects. Additionally, our method can well preserve the intra-modality and inter-modality similarities. Extensive experiments are carried out on three cross-modal datasets, and the results show that the proposed CHLS outperforms the representative cross-modal ranking methods, such as principal component analysis (PCA), LPP, CCA, semantic matching (SM) [26], semantic correlation matching (SCM) [26], and the most recent Collective Matrix Factorization Hashing (CMFH) [1]. CHLS obtains high mean average precision (MAP) due to the high-order relationship among samples captured by combinative hypergraph learning in subspace. The main contributions of this paper are summarized as follows:

(1) We formulate the cross-modal ranking as a hypergraph learning problem in the common semantic subspace. Different from most of the existing graph based methods, we explore the hypergraph learning in the common semantic subspace in cross-modal ranking scenario to capture the high-order relationship. Common subspace learning is conducted firstly to learn a common semantic space for bridging the heterogeneity gap among different modalities. By so doing, the correlation of paired samples from different modalities can be maximized in the expected subspace. To capture the high-order relationship among more than two samples, we construct a hypergraph based on the similarity matrix. Then, the ranking scores can be achieved by solving the regularizer on the hypergraph. Experimental results show the effectiveness of investigating the high-order relationship combined with the pair-wise relationship.
(2) In addition, we propose a combinative hypergraph based on the inter-modality and intra-modality similarities. By incorporating the intra-modality similarity into the inter-modality similarity, we can construct a combinative hypergraph to encode the fused similarity information, which can further enhance the cross-modal ranking performance.

The rest of this paper is organized as follows. In Section 2, we briefly overview the related work on subspace learning and cross-modal hashing methods for cross-modal retrieval and ranking tasks. In Section 3, the proposed CHLS approach is presented in detail. Section 4

reports the experimental settings and results on three representative cross-modal datasets. Finally, we conclude our work in Section 5.

2 Related work

Recently, the research of multiple views and modalities in computer vision tasks including visual tracking, object recognition, and visual retrieval. For example, [12–14] investigate the fusion of multiple views in the application of visual tracking. In terms of objective recognition, [15, 16] propose the multiple views fusion at source and feature level to perform palmprint recognition. Here we mainly focus on the connected work on multi-modal visual retrieval especially on cross-modal retrieval. Since cross-modal retrieval plays an important role in various applications, most approaches focused on cross-modal retrieval, but few considered results ranking. In the following overview, we summarize both representative approaches in cross-modal retrieval and ranking. Most of the proposed methods can be categorized into two groups i.e. subspace learning and cross-modal hashing.

2.1 Methods based on subspace leaning

The main challenge in cross-modal retrieval and ranking tasks is how to measure the similarity among different modalities due to the heterogeneity gap across them. To address this challenge, a number of studies have been proposed to learn a common subspace. Through mapping the original modalities into a common subspace, the semantic gap can be bridged and the similarity among different modalities can be measured directly by the distance metric. There are a lot of classicial methods for subspace learning, such as PCA, LDA, LPP, and 2D random projection [17, 20]. One of the most popular methods in subspace learning for cross-modal retrieval is CCA [26]. The correlation of projected modalities in the common subspace can be maximized through CCA. Thus, CCA can measure the similarity between different modalities, and also can preserve the maximized pair-wise relationships across modalities, i.e. pair-wise inter-similarities. Another widely used method is PLS [30]. PLS and CCA both learn transformations to map different original modalities into a latent common subspace in which the similarity can be computed directly. However, PLS differs from CCA in that PLS is a regression model which projects data from one modality to another [6]. Wang [31] et al. proposed a graph model which utilizes the content and semantics similarities as well as the interaction between different modalities for cross-modal retrieval. Our method mainly differs it in the high-order relationship capture. In [29], deep canonical correlation analysis with progressive and hypergraph learning is used for learning the common subspace and performing cross-modal retrieval. In [37], l_{21}-norms and graph regularization are coupled with a linear regression to learn projection matrices for mapping different modal data into the common space. With the coupled items, it can preserve the inter-modality and intra-modality similarities. In order to find the common structure hidden in different modalities, a compound regularization framework was proposed to address pairwise constraint [3]. Furthermore, multimodal subspace clustering was used to learn the common structure. Different from previous methods, a supervised consistent feature representation learning method was proposed in [9], with the capability of dealing with unpaired training samples. A joint graph regularized multimodal subspace learning approach was proposed in [39] to better explore the cross-modal correlation and the local manifold structure. The difference between them and the proposed method is that we construct a combinative hypergraph and the main focus is

ranking cross-modal query results. In these literatures, a graph regularization is widely used to preserve the inter-similarities among multimodal features. Unlike these, the proposed CHSL aims at capturing the high-order rather than pair-wise relationship. Moreover, a combinative hypergraph is constructed to encode the inter-modality and intra-modality similarities.

2.2 Methods based on cross-modal hashing

Cross-modal hashing is another widely used method for cross-modal retrieval and ranking. Hash functions are learned to map the original features to a Hamming space. The cross-modal hashing methods firstly project different modalities into a common Hamming space with minimizing an loss function. Then, the compact hash codes of cross modalities are obtained according to the learned linear or nonlinear hashing functions. It can address the large-scale cross-modal retrieval effectively and efficiently. Most of the cross-modal hashing researches differ in designing different loss functions. Collective Matrix Factorization Hashing (CMFH) [1] learns a common latent semantic space associated with linear projections for different modalities by factorizing data matrices jointly under the constraints of common factor. Cross-View Hashing (CVH) [11] formulates the problem of learning hashing functions as a generalized eigenvalue problem. Supervised Matrix Factorization Hashing [22] was proposed to seamlessly integrate semantic labels into the hashing leaning procedure for large-scale data modeling. Joint coupled-hashing [23] was proposed to firstly learn a embedding for each modality, and then Hamming space is learned through the embedding with another modality. The weakness of it is that it ignores the common semantics and the high-order relationship among samples. In [47], the Cross-Modal Self-Taught Hashing (CMSTH) was proposed. In CMSTH, unlabeled data are also used for training to obtain a better semantic correlation. In [22], a novel cross-modality hashing algorithm termed Supervised Matrix Factorization Hashing (SMFH) was proposed to tackle the multi-modal hashing problem where a collective non-negative matrix factorization across various modalities is performed. Alternating Co-Quantization (ACQ) was proposed to minimize the binary quantization errors in cross-modal hashing [7]. However, most of the existing cross-modal hashing methods considered the pair-wise relationships merely. They cannot capture the high-order relationship among more than two instances. Although the proposed CHLS in this paper is a real-valued approach based on subspace learning using CCA, it captures the high-order relationship that can boost the performance of cross-modal ranking.

Besides the subspace and cross-modal hashing based methods, a variety of approaches are also presented for the cross-modal retrieval problem, such as deep learning based method [6], graph-based method [45], multi-view method [8], and dictionary based method [49], etc. In recently, deep learning has drawn considerable interests due to its effectiveness in feature learning. A slice of methods were proposed based on deep learning [6, 24, 35]. Through deep learning, the semantic representation can be extracted effectively from the original modality. Thus, the semantic relationship across different modalities can be measured with the learned high level semantic representation. To make the cross-modal similarity computable, He et al. [6] proposed a deep and bidirectional representation learning model. Deep neural network is used to extract the semantic representation from both raw image and text data. Images and texts are mapped to a common space by passing the networks. CNN and WCNN were used to learn representations for images and texts, respectively. In [53], a linear cross-modal hashing method was proposed. It uses k-means clustering to generate a new k-dimensional representation for each sample. This method can perform cross-modal hashing in a linear complexity. Most of the existing methods can only handle offline cross-modal

subspace learning or hashing. In order to balance performance and computational complexity, [45] presented an online cross-modal hashing method, which can reduce the complexity of hash function learning significantly.

Through the overview of previous cross-modal retrieval and ranking methods, we can find that most of the existing methods fail to capture the high-order relationship among samples. Furthermore, the similarity preserving should also be considered to enhance the performance of cross-modal ranking. To this end, the motivations of our work include capturing the high-order relationship and preserving inter-modality and intra-modality similarities.

3 Combinative hypergraph learning in subspace

To bridge the heterogeneity gap among different modalities, subspace learning is employed as the first step in CHLS. Then combinative hypergraph learning is performed on the projected multimodal data, which takes both intra-modality and inter-modality similarities into consideration. Finally, we can obtain the cross-modal ranking scores that indicate the relevance of retrieved samples from another modality. The procedure of CHLS is illustrated in Fig. 1. As shown in this figure, Steps (b), (c), (d), and (e) will be analyzed in this section in detail. Feature extraction will be presented in the Section 4 associated with the description of experimental datasets. The important notations used in this paper are shown in Table 1.

3.1 Common subspace leaning

The task of cross-modal ranking requires the similarity information among different modalities. However, the similarity cannot be directly measured due to the heterogeneity gap. Hence, subspace learning methods have been proposed to learn a latent common subspace for bridging such heterogeneity gap. Thus, the similarities among various modalities can be computed directly. In this paper, a common subspace of two modalities (e.g., image and text) is learned firstly. Here, canonical correlation analysis (CCA) is utilized for common space learning. Given a pair of samples, CCA yields the maximal correlation between them. Please refer to [41] for more details on CCA.

Formally, given a set of pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$, $\mathbf{x}_i \in \Re^{d_1}$, $\mathbf{y}_i \in \Re^{d_2}$, $d_1 \neq d_2$, we cannot compute the similarity between $\mathbf{x}_i$ and $\mathbf{y}_i$ directly. CCA aims to find projections for two

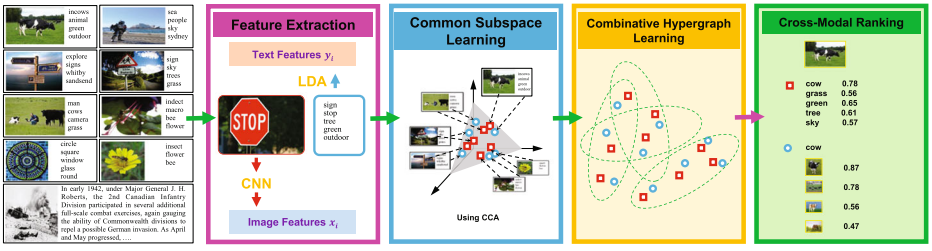


Fig. 1 The procedure of the proposed CHLS for cross-modal ranking. **a** the cross-modal multimedia datasets, **b** feature extraction from the images and texts, **c** cross-modal common subspace learning using CCA, **d** combinative hypergraph learning, and **e** cross-modal ranking

Table 1 Important notations used in this paper

Notation	Definition
$\{\mathbf{x}_i, \mathbf{y}_i\}$	The i-th cross-modal pair
n	The size of cross-modal pairs
d_1, d_2	The dimensions of given modality $\{\mathbf{x}_i, \mathbf{y}_i\}$
d	The dimension of common subspace
k	The number of k-nearest neighbors
$S_{\mathbf{xy}}$	Inter-modality similarity
$S_{\mathbf{xx}}$	Intra-modality similarity
D_v	The diagonal matrix whose diagonal entries are the degrees of vertices
D_e	The diagonal matrix whose diagonal entries are the degrees of hyperedges
W	The diagonal matrix of the hyperedge weights
H	The incidence matrix
Δ	The hypergraph Laplacian
μ	Tradeoff parameter
q	Initial ranking scores
f	The ranking scores vector
θ	Modality balance parameter

vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ that can obtain the maximal correlation. The projected data can be denoted as $\mathbf{W}_\mathbf{x}^T\mathbf{x}$ and $\mathbf{W}_\mathbf{y}^T\mathbf{y}$, and the correlation can be stated as:

$$\rho = \frac{\mathbf{W}_\mathbf{x}^T \Sigma_{\mathbf{xy}} \mathbf{W}_\mathbf{y}}{\sqrt{\mathbf{W}_\mathbf{x}^T \Sigma_{\mathbf{xx}} \mathbf{W}_\mathbf{x} \mathbf{W}_\mathbf{y}^T \Sigma_{\mathbf{yy}} \mathbf{W}_\mathbf{y}}} \tag{1}$$

where $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$ are the within-sets covariance matrices, $\Sigma_{\mathbf{xy}} = \Sigma_{\mathbf{yx}}$ are the between-sets covariance matrices, and $\mathbf{W}_\mathbf{x}$, and $\mathbf{W}_\mathbf{y}$ are the projection matrices, which can be obtained by optimizing the following maximization problem:

$$\begin{aligned} &\max_{\mathbf{W}_\mathbf{x}, \mathbf{W}_\mathbf{y}} \ \mathbf{W}_\mathbf{x}^T \Sigma_{\mathbf{xy}} \mathbf{W}_\mathbf{y} \\ &s.t. \mathbf{W}_\mathbf{x}^T \Sigma_{\mathbf{xx}} \mathbf{W}_\mathbf{x} = 1; \mathbf{W}_\mathbf{y}^T \Sigma_{\mathbf{yy}} \mathbf{W}_\mathbf{y} = 1 \end{aligned} \tag{2}$$

The optimization in (2) can be transformed into a generalized eigenvalue problem. Thus, for each pair in $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, the projected representations in the subspace can be stated as $\{\mathbf{W}_\mathbf{x}^T\mathbf{x}_i, \mathbf{W}_\mathbf{y}^T\mathbf{y}_i\}_{i=1}^n$, where $\mathbf{W}_\mathbf{x}^T\mathbf{x}_i \in \Re^d$ and $\mathbf{W}_\mathbf{y}^T\mathbf{y}_i \in \Re^d$ are the subspace representations of samples $\mathbf{x}_i$ and $\mathbf{y}_i$.

For convenience, let $\mathbf{x}' = \mathbf{W}_\mathbf{x}^T\mathbf{x}$, and $\mathbf{y}' = \mathbf{W}_\mathbf{y}^T\mathbf{y}$ denote the projected data from two modalities, where $\mathbf{x}' \in \Re^{n\times d}$, and $\mathbf{y}' \in \Re^{n\times d}$. In the following step, the projected data will be used for combinative hypergraph learning.

3.2 Hypergraph learning for ranking

Most of the existing latent subspace learning methods consider only the pair-wise relationship between two samples, but ignore the high-order relationship. The high-order term

describes relationship among more than two samples. For instance, the pair-wise similarity shows two close objects, while high-order relationship gives three or more close objects. Obviously, modeling the high-order relationship among objects can improve the ranking performance significantly, and also can return the most relevant samples to image or text query.

Let $G = (V, E, w)$ represent a hypergraph with the vertex set V, hyperedge set E, and the hyperedge weight vector w. In G, a hyperedge e_i connects more than two vertices. An incidence matrix $H \in \{0, 1\}^{|V|\times|E|}$ is used to demonstrate a hypergraph, where the entry $H(v, e) = 1$ if $v \in e$, and $H(v, e) = 0$ otherwise. Based on H, the degree $d(v)$ of a vertex $v \in V$ and the degree $\delta(e)$ of a hyperedge $e \in E$ are defined as $d(v) = \Sigma_{e\in E} w(e)H(v, e)$ and $\delta(e) = \Sigma_{v\in V} H(v, e)$, respectively [48]. Let D_v and D_e denote the diagonal matrices whose diagonal entries are the degrees of vertices and hyperedges, respectively. We also define W as a diagonal matrix whose diagonal entries are the hyperedge weights.

Given a hypergraph, the ranking scores can be obtained by optimizing the following objective function [48]:

$$\Omega(f) = \frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}\frac{w(e)H(u,e)H(v,e)}{\delta(e)} \times \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}}\right)^2 + \mu\sum_{u\in V}(f(u)-q(u))^2 \tag{3}$$

where μ is the tradeoff parameter, q is the initial ranking scores, and f is a vector denoting the final ranking scores. We can define $\Theta = D_v^{-1/2} H W D_e^{-1/2} H^T D_v^{-1/2}$, then the hypergraph Laplacian can be represented as $\Delta = I - \Theta$. Finally, the normalized objective function can be rewritten as:

$$\Omega(f) = f^T \Delta f + \mu(f-q)^T(f-q) \tag{4}$$

By differentiating $\Omega(f)$ with respect to f, the ranking scores can be computed as follows:

$$f = \left(\frac{\mu}{\mu+1}\right)\left(I - \frac{1}{\mu+1}\Theta\right)^{-1} q \tag{5}$$

3.3 The proposed CHLS

In this subsection, we present combinative hypergraph learning in subspace for cross-modal ranking. In Section 3.2, we have introduced hypergraph learning for computing ranking scores. Hence, we need to construct a hypergraph for learning cross-modal ranking scores and capturing high-order relationship among samples.

After the projection of CCA, the projected data $\mathbf{x}'$, $\mathbf{y}'$ of $\mathbf{x}$, $\mathbf{y}$ are obtained. To improve the performance of cross-modal ranking, we take both intra-modality similarity and inter-modality similarity into consideration. These two types of similarities are used to construct a combinative hypergraph. Therefore, the constructed hypergraph carries both cross-modal and within-modal information.

For example, given a paragraph of text description on 'cat', we aim to search the most relevant 'cat' images. This can be achieved by inter-modality similarity information. However, it is not enough to consider the inter-modality similarity only while ignoring the similarity information inside the queried image database. If we complement the images that are close to the returned images from cross-modal retrieval to the final image set, we can obtain complementary and more relevant 'cat' images.

From the hypergraph learning cost function in (3), we can see that the hyperedge weight vector w should be initialized with reasonable value according to the rules utilized in [51]. The similarity matrix of intra-modality is computed as follows:

$$S_{\mathbf{xx}}(i,j) = \begin{cases} \exp\left(-\frac{\|v_i - v_j\|^2}{\sigma^2}\right), & \text{if } i \neq j \\ 0. & \text{else} \end{cases} \tag{6}$$

where $v_i, v_j \in \mathbf{x}', \sigma$ is the median distance of all vertices. $S_{\mathbf{xx}}$ is the similarity information in modality $\mathbf{x}'$. Similarly, $S_{\mathbf{yy}}$ denotes the similarity information in modality $\mathbf{y}'$ with $v_i, v_j \in \mathbf{y}'$. For the correlation between $\mathbf{x}'$ and $\mathbf{y}'$, we define $S_{\mathbf{xy}} = S_{\mathbf{yx}}^T$ as the cross-modality similarity under the condition of $v_i \in \mathbf{x}', v_j \in \mathbf{y}'$.

To construct a combinative hypergraph, we utilize the cross-modal and intra-modal similarity matrices $S_{\mathbf{xx}}$, $S_{\mathbf{yy}}$ and $S_{\mathbf{xy}}$. Given a query from modality $\mathbf{x}$, the combinative similarity matrix can be obtained as:

$$S_{\mathbf{x}} = \theta * S_{\mathbf{yy}} + (1-\theta) * S_{\mathbf{xy}} \tag{7}$$

where θ represents the importance of modality $\mathbf{y}$, which is used to balance the contribution of inter-modality similarity and intra-modality similarity in constructing combinative similarity matrix. Similarly, given a query from modality $\mathbf{y}$, the combinative similarity matrix is computed as:

$$S_{\mathbf{y}} = \theta * S_{\mathbf{xx}} + (1-\theta) * S_{\mathbf{yx}} \tag{8}$$

In our work, each hyperedge consists of each vertex and its k nearest neighbors according to $S_{\mathbf{x}}$ or $S_{\mathbf{y}}$. Then we can derive the hyperedge weight of each edge as:

$$w(e_i) = \sum_{v_j \in e_i} S(i,j) \tag{9}$$

Given a query from the original modality $\mathbf{x}$ or $\mathbf{y}$, the cross-modal ranking scores can be calculated by (5).

Algorithm 1 CHLS (query from modality $\mathbf{x}$)

Input: $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n, \mathbf{x}_i \in \Re^{d_1}, \mathbf{y}_i \in \Re^{d_2}, d, k$
Output:

1: Common subspace learning
2: Compute $\mathbf{x}' \in \Re^{n \times d}, \mathbf{y}' \in \Re^{n \times d}$ by CCA
3: Compute intra-modality similarity $S_{\mathbf{yy}}$ and inter-modality similarity $S_{\mathbf{xy}}$
4: Compute combinative similarity matrix by $S_{\mathbf{x}} = \theta * S_{\mathbf{yy}} + (1-\theta) * S_{\mathbf{xy}}$
5: **for** each vertex in $S_{\mathbf{x}}$ **do**
6: collect its k-nearest neighbors, and generate a hyperedge
7: **end for**
8: Compute H, W, D_e, D_v, and $\Theta = D_v^{-1/2} H W D_e^{-1/2} H^T D_v^{-1/2}$.
9: Given a query from modality $\mathbf{x}$, compute ranking scores by $f = \left(\frac{\mu}{\mu+1}\right)\left(I - \frac{1}{\mu+1}\Theta\right)^{-1} y$
10: **return** f

3.4 Algorithm and implementation

The whole algorithm is summarized in Algorithm 1. Common subspace learning is performed in Steps 1 and 2. We use CCA to project the original cross-modal data into a

common latent subspace, where the dimensionalities of different modalities are equal, which results in a directly similarity measurement as described in Step 3. In order to construct the combinative hypergraph, we first conduct a similarity fusion as Step 4. Then, in the combinative hypergraph learning, we take the inter-modality similarity and intra-modality similarity into consideration by computing the combinative similarity matrix with a balancing parameter θ. Subsequently, in Steps 5-7, a hypergraph is constructed based on the combinative affinity matrix computed in Step 4. In this paper, we use k-nearest neighbors to generate the hyperedges, which is a widely used method in the construction of hypergraph. Finally, cross-modal ranking can be performed by given queries as illustrated in Step 8.

3.5 Complexity analysis

The computational cost of the proposed CHLS model consists of three parts: 1) common subspace learning, 2) combinative hypergraph learning, and 3) ranking. In the common subspace learning phase, the computational cost of CCA is $O\left(n\eta^2 + \eta^3\right)$ [27], where $\eta = \max(d_1, d_2)$, $O\left(n\eta^2\right)$ denotes the cost of computing the covariance matrices, and $O\left(\eta^3\right)$ represents the cost of matrix multiplication, inverse and eigenvalue decomposition. We can see that, the complexity depends on the maximal dimensionality of input modalities. After the preparation of original data such as feature extraction and dimensionality reduction, the cost of common subspace learning will be reduced due to a small η. The complexity of combinative hypergraph learning is $O\left(n^2\right)$. From the procedure in Algorithm 1, the time complexity of CHLS is dominated by the cross-modal ranking i.e. the problem in (5) which leads to a cost of $O\left(n^3\right)$. Fortunately, we can adopt an iterative method to solve (5) which can reduce the computational cost to $O\left(n^2\right)$ [2].

3.6 Extension to out-of-sample problem

Although we mainly focus on cross-modal ranking in this paper, many extensions of this basic idea are possible. Cross-modal ranking re-ranks the relevance scores of the correlated samples from another modality in the database. It also can be extended to the cross-modal retrieval with the new query which is not in the database. Firstly, we can find the most similar sample in the database according to intra-modality similarity. We then use the selected sample as a query to perform a cross-modal ranking. Finally, it is straightforward to adapt Algorithm 1 presented above to solve the new cross-modal retrieval problem, and the ranking results can be returned to an out-of-sample query easily. The extension to out-of-sample instances is illustrated in Algorithm 2.

Algorithm 2 Extension to out-of-sample (query from modality $\mathbf{x}$)

Input: out-of-sample $\mathbf{x}_{out}$, $\mathbf{x}$, $\mathbf{y}$
Output: Cross-modal retrieval results

1: Project out-of-sample query $\mathbf{x}_{out}$ into the common subspace using CCA
2: Find $\mathbf{x}_i{}'$ using KL distance $w.r.t \arg\min_i(KL(\mathbf{x}_i{}', \mathbf{x}_{out}))$; $i = 1, 2, ..., n$
3: Adapt Step 8 in Algorithm 1 with the query $\mathbf{x}_i{}'$
4: **return** cross-modal retrieval results according to ranking scores f

4 Experiments and results analysis

In this section, we carry out experiments to evaluate the proposed method in this paper. The three benchmark datasets used in the experiments are introduced firstly. Next, the evaluation metrics and implementation settings are described. We then present and analyze the performance of all the methods. Finally, the parameter sensitivity is further investigated.

4.1 Datasets

The Wiki image-text dataset [26], which contains 2866 pairs of image and text is utilized to evaluate the effectiveness of the proposed CHLS. These pairs are classified into 10 classes. We randomly select 2150 pairs for training, and the remained 716 for testing. For each sample in the image modality, we use the Convolutional Neural Networks [33] to extract a 4096-dimensional feature vector. Then principal component analysis (PCA) is performed on the 4096-dimensional vector to remove the redundancy, which leads to a compact 128-dimensional feature vector as the representation of each image. For each instance in the text modality, LDA model and PCA are applied for learning a 100-dimensional representation.

The MIR Flickr dataset consists of 25000 images along with the assigned tags. We prune the original MIR Flickr by selecting the images annotated by at least one tag, which leads to a new dataset with 24 classes. We randomly take 12054 samples for training, and the remained 8460 instances for testing. Similarly, the image modality representation is a 128-dimensional vector, and the text modality representation is a 100-dimensional topic feature vector.

The Pascal VOC dataset contains 1000 image-text pairs, which can be categorized into 20 different categories. We randomly select 400 pairs of image-text to construct training set, and the rest 600 pairs for testing set. The representation of each modality is similar to Wiki image-text and MIR Flickr as mentioned above.

4.2 Evaluation metrics

Mean average precision (MAP) is employed as the performance measurement. MAP is the mean of average precision (AP). Additionally, we evaluate the precision and recall through the precision-scope and recall-scope curves, which can reveal the performance of cross-modal ranking remarkably.

4.3 Experimental settings

We compare the proposed CHLS against several representative state-of-the-art methods, such as CCA [26], PCA [44], LPP [5], SM [26] and SCM [26] for cross-modal ranking in terms of Image query and Text query. Three distance metrics, normalized correlation (NC), L2 distance, and Kullback-Leibler divergence (KL) are used in the first set of experiments. The computational formulas are stated in (10–12), where $\mathbf{x}$ and $\mathbf{y}$ are two vectors. Additionally, we compare our method against the cross-modal hashing method CMFH [1], in which the hamming distance (HD) is used to measure similarity. Then, the one with the best performance in the first set of experiments is used in the other experiments. The parameters in the proposed CHLS are set empirically. In detail, the k-nearest parameter k is set to 5, μ is set to 0.9, and the modality importance parameter θ is set to 0.2. Since the dimension of subspace cannot be fixed for various query tasks and distance metrics, d is set to the best

empirical value. In addition, the whole experiments results in this paper are obtained on the testing set. That is to say, we perform cross-modal retrieval and ranking in the testing set to validate the effectiveness of the proposed method compared against other approaches.

$$D_{NC}(\mathbf{x}, \mathbf{y}) = \frac{-\mathbf{x}\mathbf{y}^T}{\|\mathbf{x}\| \, \|\mathbf{y}\|} \tag{10}$$

$$D_{L2}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 \tag{11}$$

$$D_{KL}(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{x}(i) \log \frac{\mathbf{y}(i)}{\mathbf{x}(i)} \tag{12}$$

4.4 Performance of cross-modal ranking

4.4.1 Results on WiKi dataset

Table 2 reports the MAP scores of the compared methods and the proposed CHLS with different distance metrics on the Wiki dataset. As shown in Table 2, the proposed CHLS significantly outperforms the competitors that thanks to the computation of cross-modal ranking scores in the common subspace which bridges heterogeneity gap. CHLS performs better than CCA, SM and SCM with the consideration of both intra-similarity and inter-similarity information. Another reason is that our CHLS captures the high-order relationship among more than two samples rather than just pair-wise relationship. In general, CCA, SM, and SCM are superior to PCA and LPP. The reason is that different from PCA and LPP, CCA, SM, and SCM obtain the maximized cross-modal correlation between paired sam-

Table 2 Performance comparison of MAP on the Wiki dataset

Methods	Distance metric	Image query	Text query
CCA [26]	NC	0.3245	0.2841
	L2	0.2886	0.2483
	KL	0.1672	0.1509
PCA [44]	NC	0.1451	0.1140
	L2	0.1433	0.1287
	KL	0.1237	0.1186
LPP [5]	NC	0.1414	0.1178
	L2	0.1426	0.1219
	KL	0.1216	0.1155
SM [26]	NC	0.4594	0.3968
	L2	0.3958	0.3909
	KL	0.3677	0.3768
SCM [26]	NC	0.3603	0.3291
	L2	0.3225	0.3224
	KL	0.3000	0.3149
CMFH [1]	HD	0.2183	0.2278
CHLS	NC	0.1601	0.1893
	L2	0.3458	0.3683
	KL	0.6447	0.5941

ples, which is more effective for reducing the heterogeneity gap. CMFH also achieves the better performance than PCA and LPP, due to the common semantics learning. In addition, we can observe that the proposed CHLS obtains the best performance with the KL distance metric. Thus, in the following experiments, the KL distance metric is used to measure similarity.

The precision-scope (a-b) and recall-scope (c-d) curves of different methods on the Wiki dataset are shown in Fig. 2, in which the precision value is obtained based on the best performance parameter setting in Table 2. The scope is specified by the number (scope = 10 to 710) of top-ranked samples returned to retrieval. As shown in Fig. 2a–b, the proposed CHLS achieves better performance on precision of both image query and text query, indicating that the top-scope samples in cross-modal ranking results of CHLS are more relevant to the given query sample. Additionally, we can see that for both image query and text query, the proposed CHLS recalls 80% relevant instances in a small scope, while its counterparts recall 80% relevant instances with much larger scopes. It demonstrates that the proposed CHLS significantly outperforms the other methods in cross-modal ranking. This is because CHLS not only performs cross-modal subspace learning, but also incorporates the intra-modality and inter-modality similarities into the learning of combinative hypergraph. Furthermore, hypergraph captures the high-order relationship among more than two instances, which further enhances the cross-modal ranking performance.

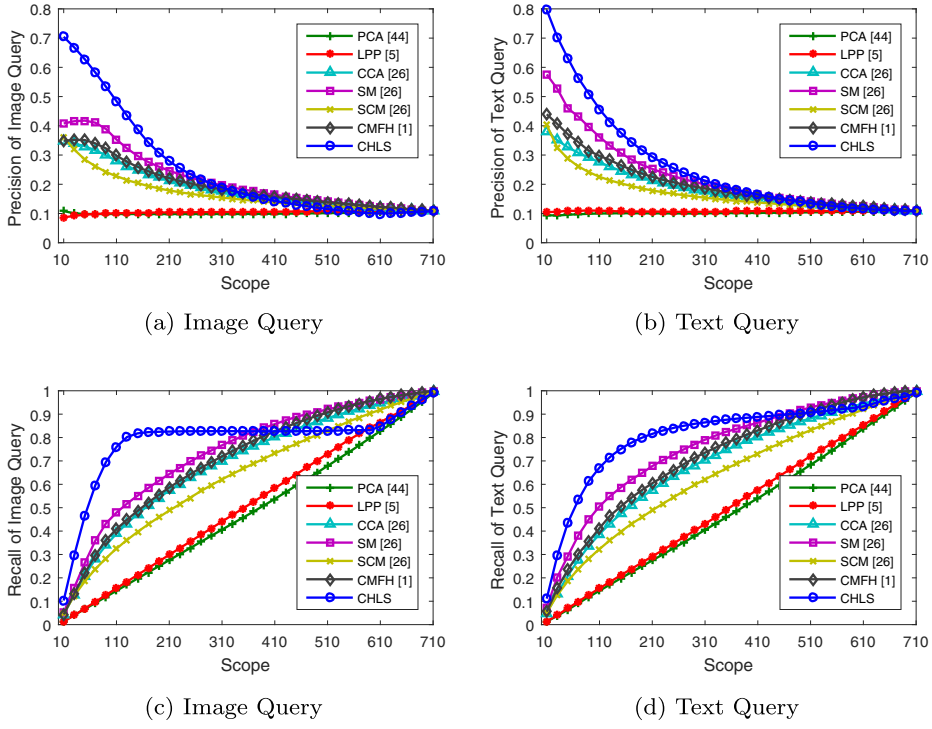


Fig. 2 Precision-scope curves (**a**–**b**) and recall-scope curves (**c**–**d**) of the proposed CHLS and its competitors on the Wiki dataset for both cross-modal ranking tasks i.e. Image query and Text query with scope = 10 to 710

4.4.2 Results on MIR Flickr dataset

The MAP scores obtained by PCA, LPP, CCA, SM, SCM, CMFH, and CHLS on the MIR Flickr dataset are shown in Table 3. It can be seen that, the proposed CHLS substantially outperforms the other counterparts. The results similar to those on the Wiki dataset are achieved. We also can observe that, PCA and LPP still obtain the poor performance on cross-modal ranking. It demonstrates that, reducing the dimensionality of different modalities to a same value is not beneficial to bridging the heterogeneity gap. The reason is that they ignore the correlations of different modalities. In contrast, CCA, SCM, SM, and CHLS that learn a common subspace by encoding the cross-modal correlation into it achieve consistent better performance than PCA and LPP in cross-modal ranking. The proposed CHLS and SCM perform better than CCA, which shows the effectiveness of exploiting additional cross-modal correlation in the latent subspace learned by CCA. Furthermore, the proposed CHLS achieves the best performance compared against SM, SCM, and CMFH, demonstrating that the method proposed in this paper by constructing a combinative hypergraph which takes inter-modality similarities and intra-modality similarities into consideration is beneficial for cross-modal ranking.

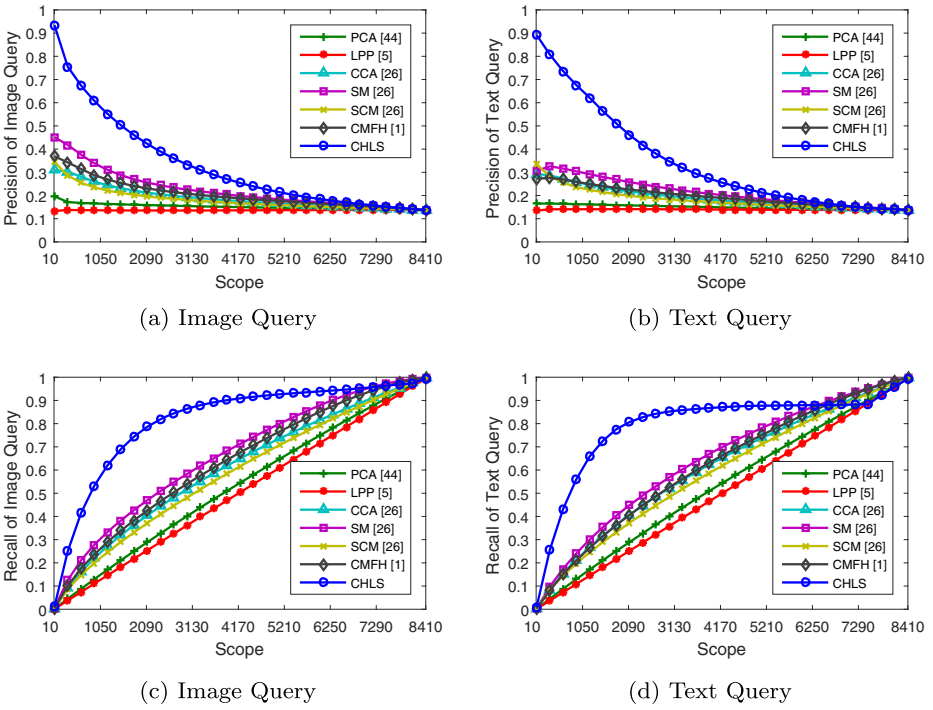
The corresponding precision-scope (a-b) and recall-scope (c-d) curves for cross-modal ranking tasks i.e. image query and text query are plotted in Fig. 3. The precision and recall are obtained with varied scope from 10 to 8410. We can see that, the proposed CHLS consistently achieves the better performance than PCA, LPP, CCA, SM, SCM, and CMFH for both precision and recall in the cross-modal ranking tasks. As shown in Fig. 3c–d, we can see that the recall of the proposed CHLS increases dramatically, reaching 80% at a small scope 2300 (27% of dataset scale) for both cross-modal ranking tasks i.e. image query and text query, while PCA, LPP, CCA, SM, SCM, and CMFH increase gradually. Thus, our proposed CHLS can return more relevant cross-modal ranking instances within a small top scope that can be beneficial to multimedia retrieval across multiple modalities. Specifically, users of commercial browsers who mainly focus on the top 10 or 20 results would prefer to use the proposed CHLS for cross-modal multimedia retrieval with high MAP score and recall.

4.4.3 Results on Pascal VOC dataset

The MAP scores of PCA, LPP, CCA, SM, SCM and CHLS on the Pascal VOC dataset are reported in Table 4. The proposed CHLS achieving MAP scores of 0.4303 and 0.4378 for image query and text query, respectively, is superior to PCA (MAP scores of 0.0802 and 0.0659), and LPP (MAP scores of 0.0714 and 0.0676) by exploiting the correlation

Table 3 Performance comparison of various approaches on the MIR Flickr dataset

Methods	Image query	Text query	Average
PCA [44]	0.1624	0.1662	0.1643
LPP [5]	0.1436	0.1501	0.1469
CCA [26]	0.2440	0.2483	0.2462
SM [26]	0.3218	0.3091	0.3155
SCM [26]	0.2724	0.2598	0.2661
CMFH [1]	0.2183	0.2278	0.2231
CHLS	0.6501	0.7381	0.6941



(a) Image Query (b) Text Query

(c) Image Query (d) Text Query

Fig. 3 Precision-scope curves (**a**–**b**) and recall-scope curves (**c**–**d**) of the proposed CHLS and its competitors on the MIR Flickr dataset for both cross-modal ranking tasks i.e. Image query and Text query with scope = 10 to 8410

of different modalities. CHLS also outperforms CCA (MAP scores of 0.2585 and 0.2437), which demonstrates the effectiveness of encoding high-order relationship into the cross-modal correlation by learning a hypergraph in the common subspace projected by CCA. CHLS not only captures the high-order relationship, but also preserves the inter-modality and intra-modality similarities by constructing a combinative hypergraph. Therefore, we can see that CHLS achieves better MAP scores compared to SCM. As reported in Table 4, the MAP score of SM for image query is slightly better than our CHLS. However, for text query, our CHLS outperforms all the other competitors, and obtains the comparable average

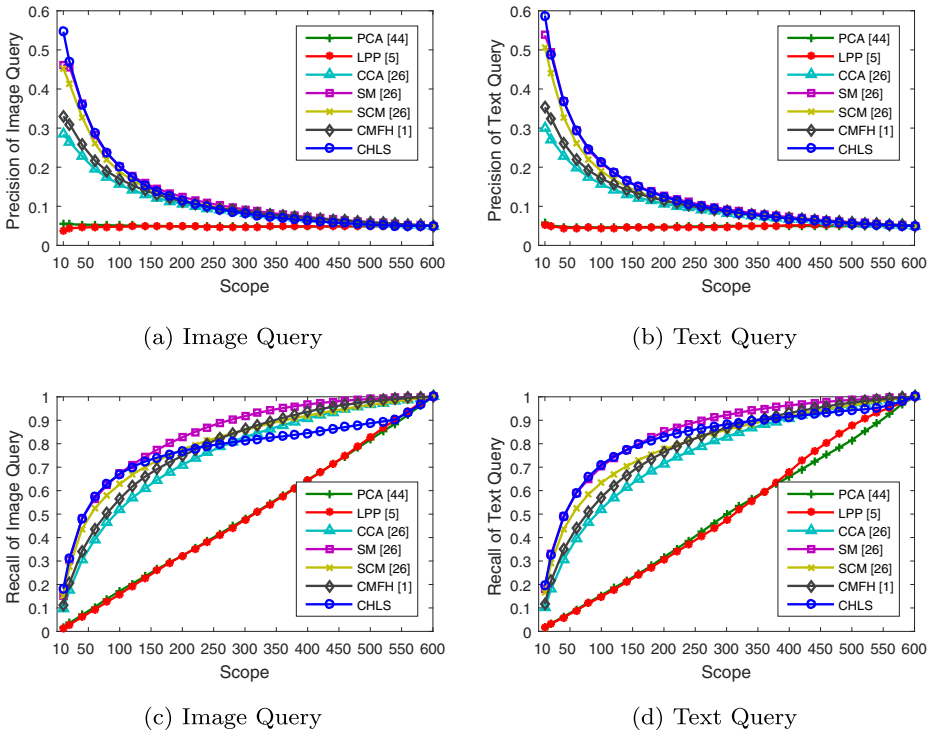
Table 4 Performance comparison of various approaches on the Pascal VOC dataset

Methods	Image Query	Text Query	Average
PCA [44]	0.0802	0.0659	0.0731
LPP [5]	0.0714	0.0676	0.0695
CCA [26]	0.2585	0.2437	0.2511
SM [26]	0.4440	0.4273	0.4357
SCM [26]	0.3105	0.2895	0.3000
CMFH [1]	0.3821	0.3798	0.3810
CHLS	0.4303	0.4378	0.4341

MAP to SM. This may because the scale of testing data is small. In addition, our CHLS still outperforms the most recent cross-modal hashing method CMFH.

The precision-scope (a-b) and recall-scope (c-d) curves of different approaches on the Pascal VOC are shown in Fig. 4. From the precision-scope curves in Fig. 4a–b, we can see that the precision of our proposed CHLS is slightly superior to that of SM, and both of them outperform the other methods. The trend of recall curves of the proposed CHLS is similar to SM. However, the recall of CHLS is slightly worse than SM when the scope is greater than 150. This is because the proposed CHLS mainly uses the k-nearest neighbors to construct hyperedges, thus very few samples will be regarded as irrelevant ones. Therefore, from Figs. 2c–d, 3c–d and 4c–d, we can see that the proposed CHLS obtains a desirable high recall at an early stage with a small scope and then keep steady for a long scope. Although our CHLS cannot always outperforms the compared methods regarding to recall-scope curves, however, it recalls the 80% relevant samples at a much earlier stage than others.

Therefore, we can conclude that preserving the correlation of different modalities in the latent subspace will be beneficial to bridge the heterogeneity gap. Second, exploiting high-order relationship among ranked instances by hypergraph learning contributes to cross-modal ranking. Third, it is useful for integrating intra-modality and inter-modality similarities to construct a combinative hypergraph, which further improves the cross-modal ranking performance.



(a) Image Query (b) Text Query

(c) Image Query (d) Text Query

Fig. 4 Precision-scope curves (**a–b**) and recall-scope curves (**c–d**) of the proposed CHLS and its competitors on the Pascal VOC dataset for both cross-modal ranking tasks i.e. Image query and Text query with the scope = 10 to 600

4.5 Parameter sensitivity

We also investigate the effects of different parameters. In the proposed model, there are three parameters should be determined, i.e., the number of nearest neighbors k, the dimension of learned common subspace d, and the tradeoff parameter θ. Here, we test the sensitivity of parameters k and d in CHLS on the Wiki dataset. Parameter k varies from 5 to 30 with a step of 5, and d increases from 10 to 100 with a step of 10. The cross-modal ranking MAPs are plotted in Fig. 5. We can observe that CHLS always obtains the highest MAP in varied subspace dimensionalities with different distance metrics when $k = 5$. The results

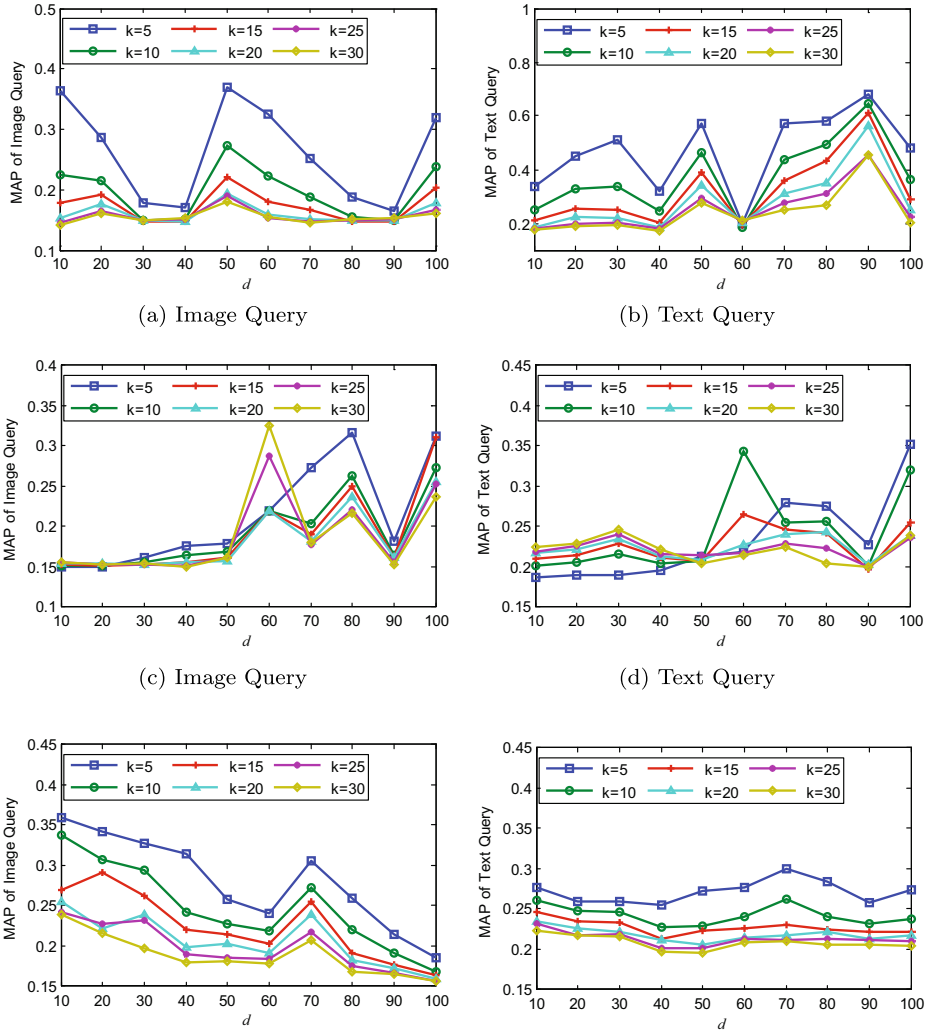


Fig. 5 Experimental results of the sensitivity of parameters k and dimension of subspace d. The left column shows results of Image Query, and the right column shows results of Text Query. From the first row to the third row, KL, NC and L2 distance metric is used, respectively

(a) Wiki (b) MIR Flickr (c) Pascal VOC

Fig. 6 Investigation of the effect of modality balance parameter θ

demonstrate that we can fix $k = 5$ in the CHLS model that can achieve promising high MAP for cross-modal ranking. This may mainly due to the fact that if k is too small, several relevant samples cannot be included into a hyperedge. If it is too large, many irrelevant sample would be included into a hyperedge which degrades the correlation among relevant samples. Hence, we can see that the performance decreases with the increasing of k.

For parameter d, we observe that the performance of our method has several fluctuations. This is because different dimensions of common subspace influence the manifold and correlation preserving and also the discriminative power of subspace representation in that common subspace. This may be caused by the subspace learning algorithm CCA, which aims at maximizing the correlation. However, the results also show that the selection of distance metric influences the optimal dimension of subspace. To the best of our knowledge, there have been few researches that tackle the problem of how to determine the optimal dimension of common subspace. Although the best performance of image query and text query are obtained with different values of d, we can always find a common d which is highly desirable for both image query and text query, such as $d = 50$ in Fig. 5a–b, $d = 100$ in Fig. 5c–d, and $d = 70$ in Fig. 5e–f. In our work, empirical values of subspace dimensionality are selected based on the distance metric.

In terms of the tradeoff parameter θ, the effects of θ on three datasets are shown in Fig. 6. Generally, we can see that it obtains inferior performance when $\theta = 0$ or $\theta = 1$. This may be caused by the consideration of only one similarity information in the combinative hypergraph learning. Hence, the results show the effectiveness of fused similarity. In addition, we can observe that the performance is insensitive to the value of θ when $0 < \theta < 1$.

5 Conclusion

In this paper, we have proposed a novel approach based on combinative hypergraph learning in subspace for cross-modal ranking. In order to bridge the heterogeneity gap of different modalities, we firstly introduced common subspace learning based on CCA, which can preserve the maximal correlation in projected common subspace. Subsequently, the combinative hypergraph learning is performed for two cross-modal ranking tasks i.e. image query and text query. Experimental results have demonstrated the effectiveness of the proposed CHLS which exploits the high-order relationship and takes both intra-modality and inter-modality similarities into consideration. We can conclude that exploiting high-order relationship by hypergraph and considering intra-modality and inter-modality similarities are beneficial to cross-modal ranking, with 80% relevant instances returned at a much earlier stage compared against state-of-the-art methods. Our future work aims at generalizing

our method to achieve not only ordinary cross-modal task but also for the scenario with unequal numbers of samples from different modalities.

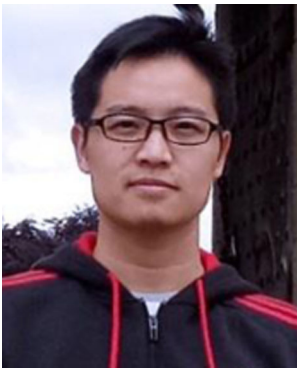
Acknowledgements This work is jointly supported by the Nature Science Foundation of China under Grant 61672123, the State Key Program of National Natural Science of China under Grant U1301253, the Science and Technology Planning Key Project of Guangdong Province under Grant 2015B010110006, the National Key Research and Development Program of China under Grant 2016YFD0800300, and the Chinese Scholarship Council.

References

1. Ding G, Guo Y, Zhou J (2014) Collective matrix factorization hashing for multimodal data. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp 2083–2090. https://doi.org/10.1109/CVPR.2014.267
2. Gao Y, Wang M, Luan H, Shen J, Yan S, Tao D (2011) Tag-based social image search with visual-text joint hypergraph learning. In: ACM international conference on Multimedia, pp 1517–1520
3. He R, Zhang M, Wang L, Ji Y, Yin Q (2015) Cross-modal subspace learning via pairwise constraints. IEEE Trans Image Process 24(12):5543–5556. https://doi.org/10.1109/TIP.2015.2466106, arXiv:1411.7798v1
4. He X (2004) Incremental semi-supervised subspace learning for image retrieval. In: MM'04, pp 2–8
5. He X, Niyogi P (2004) Locality preserving projections. Neural Inf Proces Syst 16:153–160
6. He Y, Xiang S, Kang C, Wang J, Pan C (2016) Cross-modal retrieval via deep and bidirectional representation learning. IEEE Trans Multimedia 18(7):1363–1377. https://doi.org/10.1109/TMM.2016.2558463
7. Irie G, Arai H, Taniguchi Y (2016) Alternating co-quantization for cross-modal hashing. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1886–1894. https://doi.org/10.1109/ICCV.2015.219
8. Jin Y, Cao J, Ruan Q, Wang X (2014) Cross-modality 2D-3D face recognition via multiview smooth discriminant analysis based on ELM. J Electr Comput Eng 2014(21):1–10. https://doi.org/10.1155/2014/584241
9. Kang C, Xiang S, Liao S, Xu C, Pan C (2015) Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans Multimedia 17(3):370–381. https://doi.org/10.1109/TMM.2015.2390499
10. Kitanovski I, Strezoski G, Dimitrovski I, Madjarov G, Loskovska S (2016) Multimodal medical image retrieval system. Multimedia Tools Appl 76:2955–2978. https://doi.org/10.1007/s11042-016-3261-1
11. Kumar S, Udupa R (2011) Learning hash functions for cross-view similarity search. In: Proceedings of International Joint Conference on Artificial Intelligence, Barcelona, Spain, pp 1360–1365. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-230
12. Lan X, Ma AJ, Yuen PC, Chellappa R (2015) Joint sparse representation and robust feature-level fusion for multi-cue visual tracking. IEEE Trans Image Process 24(12):5826–5841. https://doi.org/10.1109/TIP.2015.2481325
13. Lan X, Ma AJ, Yuen PC (2014) Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp 1194–1201. https://doi.org/10.1109/CVPR.2014.156
14. Lan X, Zhang S, Yuen PC (2016) Robust joint discriminative feature learning for visual tracking. In: IJCAI, pp 3403–3410
15. Leng L, Li M, Kim C, Bi X (2017) Dual-source discrimination power analysis for multi-instance contactless palmprint recognition. Multimedia Tools Appl 76(1):333–354
16. Leng L, Li M, Leng L, Teoh ABJ (2013) Conjugate 2dpalmhash code for secure palm-print-vein verification. In: 2013 6th International Congress on Image and Signal Processing (CISP), vol 03, pp 1705–1710. https://doi.org/10.1109/CISP.2013.6743951
17. Leng L, Zhang J, Chen G, Khan MK, Alghathbar K (2011) Two-directional two-dimensional random projection and its variations for face and palmprint recognition. In: International Conference on Computational Science and Its Applications, Springer, pp 458–470
18. Leng L, Zhang J, Khan MK, Chen X, Alghathbar K (2010) Dynamic weighted discrimination power analysis: a novel approach for face and palmprint recognition in dct domain. Int J Phys Sci 5(17):2543–2554

19. Leng L, Zhang J, Xu J, Khan MK, Alghathbar K (2010) Dynamic weighted discrimination power analysis in dct domain for face and palmprint recognition. In: 2010 International Conference on Information and Communication Technology Convergence (ICTC), pp 467–471. https://doi.org/10.1109/ICTC.2010.5674791
20. Leng L, Zhang S, Bi X, Khan MK (2012) Two-dimensional cancelable biometric scheme. In: 2012 International Conference on Wavelet Analysis and Pattern Recognition, pp 164–169. https://doi.org/10.1109/ICWAPR.2012.6294772
21. Lienhart R, Romberg S, Hörster E (2009) Multilayer pLSA for multimodal image retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), Santorini, GR, pp 1–8. https://doi.org/10.1145/1646396.1646408
22. Liu H, Ji R, Wu Y, Hua G (2016) Supervised matrix factorization for cross-modality hashing. In: Proceedings of International Joint Conference on Artificial Intelligence 2016-Janua(7), pp 1767–1773. https://doi.org/10.1109/TIP.2016.2564638, arXiv:1603.05572
23. Liu Y, Chen Z, Deng C, Gao X (2016) Joint coupled-hashing representation for cross-modal retrieval. In: Proceedings of the International Conference on Internet Multimedia Computing and Service, ACM, pp 35–38
24. Lu X, Wu F, Li X, Zhang Y, Lu W, Wang D, Zhuang Y (2014) Learning multimodal neural network with ranking examples. In: Proceedings of the ACM International Conference on Multimedia - MM '14, pp 985–988. https://doi.org/10.1145/2647868.2655001
25. Lu X, Wu F, Tang S, Zhang Z, He X, Zhuang Y (2013) A low rank structural large margin method for cross-modal ranking. In: Proceedings of ACM SIGIR'13, pp 433–442. https://doi.org/10.1145/2484028.2484039
26. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM International Conference on Multimedia, Firenze, Italy, pp 1–10. https://doi.org/10.1145/1873951.1873987
27. Rasiwasia N, Mahajan D, Mahadevan V, Aggarwal G (2014) Cluster canonical correlation analysis. In: Proceedings of Advances in Neural Information Processing Systems, pp 823–831
28. Rosipal R, Kr N (2006) Overview and recent advances in partial least squares. Subspace, Latent Structure and Feature Selection 3940:34–51
29. Shao J, Wang L, Zhao Z, Cai A (2016) Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval. Neurocomputing 214:618–628
30. Sharma A, Jacobs DW (2011) Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 593–600. https://doi.org/10.1109/CVPR.2011.5995350
31. Shixun W, Peng P, Yansheng L (2013) A graph model for cross-modal retrieval. In: 3rd International Conference on Multimedia Technology (ICMT-13), Atlantis Press
32. Siddiquie B, White B, Sharma A, Davis LS (2014) Multi-modal image retrieval for complex queries using. In: Proceedings of ACM International Conference on Multimedia Retrieval, Glasgow, United Kingdom, pp 1–8
33. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proceedings of International Conference on Learning Representations, pp 1–14. https://doi.org/10.1016/j.infsof.2008.09.005, arXiv:1409.1556
34. Tang J, Wang K, Shao L (2016) Supervised matrix factorization hashing for cross-modal retrieval. IEEE Trans Image Process 25(7):3157–3166. https://doi.org/10.1109/TIP.2016.2564638, arXiv:1603.05572
35. Wang C, Yang H, Meinel C (2016) A deep semantic framework for multimodal representation learning. Multimedia Tools Appl 75:9255–9276. https://doi.org/10.1007/s11042-016-3380-8
36. Wang D, Gao X, Wang X, He L (2015) Semantic topic multimodal hashing for cross-media retrieval. In: Proceedings of International Joint Conference on Artificial Intelligence, pp 3890–3896
37. Wang K, He R, Wang L, Wang W, Tan T (2016) Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans Pattern Anal Mach Intell 38(10):2010–2023. https://doi.org/10.1109/TPAMI.2015.2505311
38. Wang K, He R, Wang W, Wang L, Tan T (2013) Learning coupled feature spaces for cross-modal matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2088–2095. https://doi.org/10.1109/ICCV.2013.261
39. Wang K, Wang W, He R, Wang L, Tan T (2013) Multi-modal subspace learning with joint graph regularization for cross-modal retrieval. In: Proceedings of 2nd IAPR Asian Conference on Pattern Recognition, pp 236–240. https://doi.org/10.1109/ACPR.2013.44

40. Wang L, Sun W, Zhao Z, Su F (2017) Modeling intra- and inter-pair correlation via heterogeneous high-order preserving for cross-modal retrieval. Signal Process 131:249–260. https://doi.org/10.1016/j.sigpro.2016.08.012
41. Wang S, Gu X, Lu J, Yang J, Wang R, Yang J (2014) Unsupervised discriminant canonical correlation analysis for feature fusion. In: ICPR, pp 1550–1555. https://doi.org/10.1109/ICPR.2014.275
42. Wang S, Pan P, Lu Y, Xie L (2015) Improving cross-modal and multi-modal retrieval combining content and semantics similarities with probabilistic model. Multimedia Tools Appl 74(6):2009–2032. https://doi.org/10.1007/s11042-013-1737-9
43. Wang Y, Li P, Yao C (2014) Hypergraph canonical correlation analysis for multi-label classification. Signal Process 105:258–267. https://doi.org/10.1016/j.sigpro.2014.05.032
44. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1-3):37–52
45. Xie L, Shen J, Zhu L (2016) Online cross-modal hashing for web image retrieval. In: Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016), pp 294–300
46. Xie L, Zhu L, Chen G (2016) Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. Multimedia Tools Appl 75:9185–9204. https://doi.org/10.1007/s11042-016-3432-0
47. Xie L, Zhu L, Pan P, Lu Y (2016) Cross-modal self-taught hashing for large-scale image retrieval. Signal Process 124:81–92. https://doi.org/10.1016/j.sigpro.2015.10.010
48. Xu J, Singh V, Guan Z, Manjunath B (2012) Unified hypergraph for image ranking in a multimodal context. In: ICASSP, pp 2333–2336
49. Xu X, Yang Y, Shimada A, Ri Taniguchi, He L (2015) Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts. In: Proceedings of the ACM International Conference on Multimedia, pp 847–850. https://doi.org/10.1145/2733373.2806346
50. Yao T, Kong X, Fu H, Tian Q (2016) Semantic consistency hashing for cross-modal retrieval. Neurocomputing 193:250–259. https://doi.org/10.1016/j.neucom.2016.02.016
51. Yu J, Tao D, Wang M (2012) Adaptive hypergraph learning and its application in image classification. IEEE Trans Image Process 21(7):3262–3272
52. Zhan Y, Sun J, Niu D, Mao Q, Fan J (2015) A semi-supervised incremental learning method based on adaptive probabilistic hypergraph for video semantic detection. Multimedia Tools Appl 74(15):5513–5531. https://doi.org/10.1007/s11042-014-1866-9
53. Zhu X, Huang Z, Shen HT, Zhao X (2013) Linear cross-modal hashing for efficient multimedia search. In: Proceedings of ACM International Conference on Multimedia, Barcelona, Spain, pp 143–152. https://doi.org/10.1145/2502081.2502107
54. Zhuang Y, Wang Y, Wu F, Zhang Y, Lu W (2013) Supervised coupled dictionary learning with group structures for multi-modal retrieval. In: AAAI, pp 1070–1076

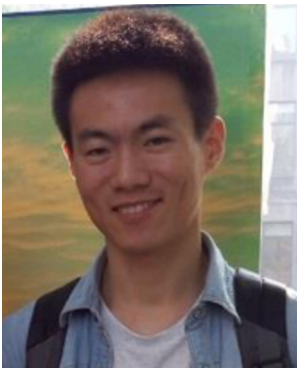


Fangming Zhong received the B.S. and M.S degree in Software Engineering from Dalian University of Technology in 2012 and 2014, respectively. He is now pursing his PhD at School of Software Technology in Dalian University of Technology, Dalian, China. His research interests include multimodal learning, cross-modal retrieval, and subspace learning.

Zhikui Chen received the B.S. degree in mathematics from Chongqing Normal University, Chongqing, China, in 1990, and the M.S. and Ph.D. degrees in mechanics from Chongqing University, Chongqing, China, in 1993 and 1998, respectively. He is currently a Full Professor with the Dalian University of Technology, Dalian, China. He is leading the Institute of Ubiquitous Network and Computing, Dalian University of Technology. His research interests are big data processing, mobile cloud computing, ubiquitous network and its computing. Dr. Chen was the General Chair of IEEE ithings2011 and IEEE Smartdata2015, the Advisor Chair of IEEE ithings2012?2015, and the Program Chair of IEEE ICDH2014.

Geyong Min is a Professor of High-Performance Computing and Networking with the Department of Mathematics and Computer Science, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, U.K. He received the Ph.D. degree in Computing Science from the University of Glasgow, Glasgow, U.K., in 2003, and the B.Sc. degree in Computer Science from the Huazhong University of Science and Technology, Wuhan, China, in 1995. His research interests include next-generation Internet, wireless communications, multimedia systems, information security, high-performance computing, ubiquitous computing, modeling, and performance engineering.

Zhaolong Ning received the M.S. and Ph. D. degrees from Northeastern University, Shenyang, China in 2014. He was a Research Fellow at Kyushu University, Japan. Currently, he is an assistant professor in School of Software, Dalian University of Technology, Dalian, China. Dr. Ning has published over 50 scientific papers in international journals and conferences. His research interests include social network, cloud computing, and scholarly big data.



Hua Zhong is now pursuing his M.S. degree in School of Software, Dalian University of Technology, Dalian, China. His research interests include cross-modal retrieval and automatic image annotation.

Yueming Hu received the Ph.D. degree in soil science from the Zhejiang Agricultural University, Hangzhou, China, in 1997, and the M.S. degree in soil science from the Northwestern Agricultural University, Xianyang, China, in 1990. He is currently a Professor with the College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China. His research interests include land resource management, geographic information system application, and agricultural information.