7-2016

Cross-Modal Self-Taught Hashing for large-scale image retrieval

Liang XIE
Wuhan University of Technology

Lei ZHU
Singapore Management University, lzhu@smu.edu.sg

Peng PAN
Huazhong University of Science and Technology

Yansheng LU
Huazhong University of Science and Technology

Published in Signal Processing, 2016 July, Volume 124, Pages 81-92
http://doi.org/10.1016/j.sigpro.2015.10.010

Cross-Modal Self-Taught Hashing for large-scale image retrieval

Liang Xie [a], Lei Zhu [b,*], Peng Pan [c], Yansheng Lu [c]

[a] *School of Science, Wuhan University of Technology, Wuhan, China*
[b] *School of Information Systems, Singapore Management University, Singapore*
[c] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China*

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form
26 September 2015
Accepted 13 October 2015
Available online 23 October 2015

Keywords:
Image retreival
Cross-modal hashing
Self-taught learning
Semantic correlation

ABSTRACT

Cross-modal hashing integrates the advantages of traditional cross-modal retrieval and hashing, it can solve large-scale cross-modal retrieval effectively and efficiently. However, existing cross-modal hashing methods rely on either labeled training data, or lack semantic analysis. In this paper, we propose Cross-Modal Self-Taught Hashing (CMSTH) for large-scale cross-modal and unimodal image retrieval. CMSTH can effectively capture the semantic correlation from unlabeled training data. Its learning process contains three steps: first we propose Hierarchical Multi-Modal Topic Learning (HMMTL) to detect multi-modal topics with semantic information. Then we use Robust Matrix Factorization (RMF) to transfer the multi-modal topics to hash codes which are more suited to quantization, and these codes form a unified hash space. Finally we learn hash functions to project all modalities into the unified hash space. Experimental results on two web image datasets demonstrate the effectiveness of CMSTH compared to representative cross-modal and unimodal hashing methods.

1. Introduction

In recent years, with the development of information and network technologies, there has been a massive explosion of multimedia data on the web. Large amounts of multimedia contents, especially images, are generated, shared and accessed by users on Wikipedia, Flickr, Youtube and other popular social websites. Web multimedia contents have two characteristics. On one hand, they contain various modalities, such as image, text, video and audio. On the other hand, the amount of them becomes rather huge. To meet the development trend of web multimedia, cross-modal retrieval and hashing have become two important techniques.

In traditional unimodal image retrieval, image examples are usually used as queries to search image database [1]. However, in real world, users may be not satisfied with image queries, but more comfortable to use other types of queries such as text and sound. Moreover, users may want to use images to search other types of data. Cross-modal retrieval, which has been extensively studied in the multimedia literature [2–5], is designed for the retrieval of heterogeneous data, e.g., using text query to retrieve images. In addition, cross-modal methods can even enhance the performance of unimodal retrieval by exploiting the correlation between image and other modality. Besides effectively correlating multi-modal data, efficiently indexing large-scale data is also important. Hashing is an efficient indexing approach which can be used to solve the retrieval of large-scale images [6–8]. It converts high-dimensional data into short binary codes, which preserve the similarity of data. Then fast search can be easily implemented by efficient XOR and bit-count operations.

* Corresponding author.
E-mail addresses: whutxl@hotmail.com (L. Xie), leizhu0608@gmail.com (L. Zhu), panpeng@mail.hust.edu.cn (P. Pan), lys@mail.hust.edu.cn (Y. Lu).

http://dx.doi.org/10.1016/j.sigpro.2015.10.010

Motivated by the success of hashing and cross-modal retrieval, recently several cross-modal hashing methods are proposed to integrate the advantages of them. Common cross-modal hashing methods learn a unified hash space which correlates different modalities, and then the search process can be accelerated based on hash codes. The performance of cross-modal hashing lies on effect of cross-modal correlation and quantization. However, most existing cross-modal hashing methods cannot well solve both these two points. Some cross-modal hashing methods are only binary versions of traditional cross-modal retrieval approaches [9,10]. They mainly focus on the cross-modal correlation, but pay less attention to the effectiveness of hash code quantization, which is also important for cross-modal hashing. For example, eigenvalue decomposition is widely used for cross-modal hashing, but it will cause considerable quantization loss in generating hash codes [11].

Cross-modal correlation analysis is essential for cross-modal retrieval/hashing, but it is not well solved by current methods. Cross-modal correlation describes the relationship between different modalities, there exist mainly two types of cross-modal correlations: content correlation and semantic correlation. Content correlation is usually analyzed by unsupervised cross-modal methods [12,13], it directly matches the contents of heterogeneous modalities. The advantage of content correlation is that it does not require supervised labels for the training data, and can be learned from the co-occurrence of different modalities. Semantic correlation is usually analyzed by supervised cross-modal methods [3,4], it matches different modalities according to semantic concepts or topics. For example, the photo and sound of a bird can be correlated by the concept 'bird'. The advantage of semantic correlation is that it describes the cross-modal data at a high level of abstraction, which is more close to the real world. As a result, methods based on semantic correlation usually achieve better performance.

Both existing supervised and unsupervised cross-modal methods have limitations in analyzing cross-modal correlation. Supervised methods can well capture the cross-modal correlation by semantic concepts, but they need the labeled training data which are difficult to obtain. As a result, supervised methods are not practical for real world application. Unsupervised methods can directly use unlabeled data for learning, but content correlation is not able to effectively describe cross-modal data. Therefore, the retrieval performance of unsupervised methods is usually worse than supervised methods. Semi-supervised methods can partly solve the above limitations [8], but they still need a small amount of training data to be labeled. This means that professional people should define specific semantic concepts and use them for labeling.

In this paper, we propose a novel method: Cross-Modal Self-Taught Hashing (CMSTH) for large-scale image retrieval. The core idea of CMSTH is that the semantic topics are learned automatically, then all modalities are correlated by these topics. Since semantic topics have high-level abstraction and contain much semantic information, CMSTH can better correlate different modalities than previous unsupervised methods. Besides, CMSTH is also more practical than supervised methods, it directly uses unlabeled data for learning. CMSTH consists of three learning steps. In the first step, we propose Hierarchical Multi-Modal Topic Learning (HMMTL) which can effectively learn the semantic information from multi-modal data. In the second step, we specifically consider the quantization loss, and transfer semantic topics to effective hash codes by Robust Matrix Factorization (RMF). At last, we learn efficient hash functions of all modalities which can project them into the unified hash space.

The contributions of this paper are listed as follows:

- We propose a novel framework: Cross-Modal Self-Taught Hashing (CMSTH) for large scale image retrieval. CMSTH automatically detects semantic topics, and constructs the semantic correlation of different modalities based on these semantic topics.
- We propose Hierarchical Multi-Modal Topic Learning (HMMTL) to learn semantic topics from multi-modal data. HMMTL simultaneously preserves intra-modal and inter-modal consistency, and assigns proper weights to different modalities. Thus HMMTL can learn topics with more semantic information.
- Since multi-modal topics are not suited to hashing, we use Robust Matrix Factorization (RMF) to transfer topics to hash codes. RMF can generate effective hash codes which also preserve the semantic correlation of topics.

When compared with a preliminary version [14]. We have made following improvements: (1) We have discussed more comprehensive survey of related work, especially on large-scale cross-modal retrieval. (2) We improve our method to make it more suited to large-scale data, and a new hash generation step is introduced. (3) We compare our method to more representative cross-modal hashing methods in experiments. The rest of this paper is organized as follows. Section 2 discusses the related methods. In Section 3, we describe the learning process of CMSTH. Section 4 shows the experimental results on two multi-modal image datasets. Finally the conclusions and future work are presented in Section 5.

2. Related work

2.1. Multi-modal and cross-modal learning

In recent years, multi-modal and cross-modal learning have been extensively studied [15]. Multi-modal learning, which is also named as multi-view learning, has been shown the effectiveness in image classification/annotation [16–20]. In [19], Multiview Matrix Completion (MVMC) is proposed for semi-supervised multilabel image classification. MVMC weightedly combines the matrix completion outputs of different modalities, and a cross-validation strategy is applied to effectively learn combination weights. Some multi-modal methods learn a unified space from all modalities [16,21,22] to improve the performance of image classification. In [22], the Multi-view Intact Space Learning (MISL) is proposed to integrate multiple modalities. For the learning of intact space, Cauchy loss is used to strengthen robustness to outliers.

Multi-modal and cross-modal learning can also solve the problem of cross-modal retrieval, and both supervised and unsupervised methods have been studied. Canonical Correlation Analysis (CCA) [23] and Canonical Factor Analysis (CFA) [13] are two representative unsupervised methods, and they are widely used in cross-modal retrieval [24,5]. The main idea of CCA and CFA is learning two correlated subspace for two different modalities, such as images and texts. Manifold learning has also been used for unsupervised cross-modal retrieval [2,25,26]. They exploit the neighborhood relation in multi-modal data, and learn a unified space to represent different modalities. Recently, with the development of deep learning, several deep methods are proposed for cross-modal retrieval, including Multimodal DBM [27] and Correspondence Autoencoder (Corr-AE) [28]. Most unsupervised methods cannot capture the semantic information in the multi-modal data, which limits their retrieval performance.

Supervised cross-modal methods correlate different modalities according to semantic labels, and they usually outperform unsupervised methods. In [24], semantic labels directly form a semantic space, then Logistic Regression and Support Vector Machine (SVM) are used to project different modalities into this space, and the performance is shown to be much better than CCA and CFA. Supervised ranking methods are recently adopted for cross-modal retrieval [29,30], they select training examples by semantic labels. The training examples in supervised methods should be labeled, thus they are not practical for real world application.

The main disadvantage of above cross-modal retrieval methods is their scalability, they are not so efficient for large-scale multi-modal data. Some cross-modal methods, such as CCA, can be transformed to hashing by quantizing each dimension of their subspace to binary code. However, since their subspace is not specifically designed for hashing, the quantization will cause significant information loss.

2.2. Unimodal hashing

Hashing is an efficient approach for large-scale image retrieval, current hashing methods can be generally divided into two categories: random projection based hashing and machine learning based hashing [31]. Locality Sensitive Hashing (LSH) [32] is one of the most representative random projection methods. LSH is data-independent, thus it may lead to ineffective hash codes in practice. Machine learning based hashing methods learn more reliable hash function by analyzing the contents of data. The representative machine learning based methods include Spectral Hashing (SH) [33], Kernelized Hashing [7], K-means Hashing [6], Anchor Graph Hashing [34] and Self-Taught Hashing (STH) [35]. STH also adopts the self-taught scheme, but the disadvantage of STH is that it cannot be applied to multi-modal data. Moreover, it only uses two steps which do not consider the quantization effect of codes. As a result, STH performs worse than our method even in unimodal retrieval.

2.3. Cross-modal hashing

The methods of cross-modal hashing are most related to our work. Similar to traditional cross-modal retrieval, both supervised and unsupervised cross-modal hashing have been researched in recent years. Supervised and semi-supervised hashing methods also require semantic labels [36–38,11], which makes them be not practical. Recently, large progress has been made for unsupervised cross-modal hashing. Representative unsupervised cross-modal hashing methods include Cross-View Hashing (CVH) [9], Inter-Media Hashing (IMH) [31], Composite

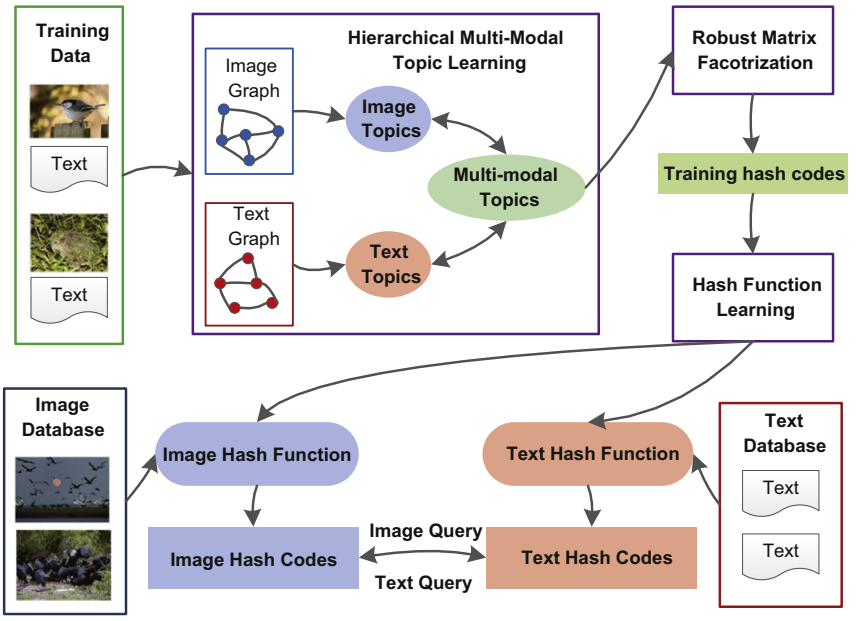

Fig. 1. The framework of CMSTH.

Hashing with Multiple Information Sources (CHMIS) [10] and Collective Matrix Factorization Hashing (CMFH) [39]. However, they only consider the content correlation of different modalities, and are unable to analyze the semantic correlation. Our method also uses unsupervised data for training, but it can effectively preserve the semantic correlation in hash codes.

3. Cross-modal self-taught hashing

The whole framework of Cross-Modal Self-Taught Hashing (CMSTH) is illustrated in Fig. 1. We can find that CMSTH uses unlabeled multi-modal data for training. CMSTH contains three steps, the first step is Hierarchical Multi-Modal Topic Learning (HMMTL), which learns semantic topics by preserving both intra-modal and inter-modal consistency. The second step uses Robust Matrix Factorization (RMF) to make our method more suited to the quantization of hash codes. Then we learn linear hash function which can efficiently project all modalities into the shared hash space. Finally, we can obtain hash codes for both image and text, and their hash codes can be directly matched by hamming distance.

3.1. Notations and definitions

Suppose there are n training multi-modal examples $E_1, \ldots, E_n$. Each example E_i contains M modalities, and $E_i = \{x_i^1, \ldots, x_i^M\}$, where x_i^m is the feature vector of mth modality. In this paper, we only consider two modalities: image and text, thus $M=2$ and E_i is an image-text pair. Images on the web are usually associated with text, and our methods can also be applied to more modalities ($M > 2$). For simplicity, a list of notations used in this paper are shown in Table 1.

3.2. Hierarchical multi-modal topic learning

The first step of CMSTH is to learn unified semantic topics in the unsupervised manner, then all modalities can be semantically correlated by these topics. Generally, combining multi-modal sources can obtain more semantic information, thus we use both image and text to learn semantic topics. Unlike traditional multi-modal methods, we use a hierarchical learning process to combine image and text. Firstly, we learn the unimodal topics which preserve the intra-modal consistency of image and text. Then we combine unimodal topics and generate the final topics which can further preserve the inter-modal consistency. The advantage of Hierarchical Multi-modal Topic Learning (HMMTL) is that it not only preserves inter-modal consistency, but also preserves intra-modal consistency. Intra-modal and inter-modal consistency are both important for cross-modal analysis [40], thus HMMTL can learn topics with more semantic information.

For each modality, we construct its intra-modal similarity graph A_m, which is defined as

$$A_{ij}^m = \begin{cases} e^{-\left\|x_i^m - x_j^m\right\|_2^2/\sigma_m}, & x_i^m \text{ and } x_j^m \text{ are } c \text{ nearest neighbors} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where σ_m is the mean of all $\left\|x_i^m - x_j^m\right\|_2^2$.

To learn the multi-modal topics from training data, we first generate unimodal topic matrix $F_m \in \mathbb{R}^{n \times t_m}$ for each modality, where t_m is the number of unimodal topics for each modality. F_m can preserve the semantic information of the mth modal features, it is obtained by minimizing the following graph Laplacian regularizer:

$$\sum_{k=1}^{t_m} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij}^m \left(\frac{f_{ik}^m}{d_{ii}^m} - \frac{f_{jk}^m}{d_{jj}^m} \right) = \mathrm{Tr}\left(F_m^T L_m F_m \right) \quad (2)$$

where f_{ik}^m is an element of F_m, and d_{ii}^m is the sum of ith row of A_m. $L_m = I - D_m^{-1/2} A_m D_m^{-1/2}$, I is the identity matrix, D_m is the diagonal matrix and its diagonal element is d_{ii}^m. $\mathrm{Tr}(\cdot)$ denotes the trace operator.

After we obtain unimodal topic matrices $F_m|_{m=1}^M$, we use them to generate the multi-modal topic matrices $F \in \mathbb{R}^{n \times t}$, where t is the number of multi-modal topics. Each F_m generates the final F by minimizing the following function:

$$\|F - F_m W_m\|_F^2 \quad (3)$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm. $W_m \in \mathbb{R}^{t_m \times t}$ is the weight matrix for the generation of F. According to (3) we can find that the intra-modal information in F_m is transferred to F, thus F also preserves the intra-modal consistency of each modality. Moreover, all F_m are combined and they should be consistent with F, thus F further preserves the inter-modal consistency.

In summary, the multi-modal topic matrix F is hierarchically generated. At first, unimodal topic matrices $F_m|_{m=1}^M$ are generated by intra-modal similarity. Then multi-modal topic matrix F is generated from all $F_m|_{m=1}^M$. We optimize the hierarchical generation in a joint learning process. By combining (2) and (3), we arrive at the following objective function:

$$\begin{aligned} \min_{F} \quad & \sum_{m=1}^{M} \left(\mathrm{Tr}\left(F_m^T L_m F_m \right) + \alpha_m^2 \|F - F_m W_m\|_F^2 \right) \\ \text{s.t.} \quad & F_m^T F_m = I, \quad m = 1, \ldots, M \\ & F^T F = I \\ & \sum_{m=1}^{M} \alpha_m = 1 \end{aligned} \quad (4)$$

where the orthogonality constraints on F and F_m are to avoid the trivial solution. α_m is the weight parameter, it reflects the importance of mth modality for the generation of F, and we can easily find that (4) is convex with respect to α_m. Generally, different modalities should have different importance for semantic learning. For example, texts usually contain more semantic information than images.

By setting the derivative of (4) w.r.t. W_m to zero, we have:

$$W_m = F_m^T F \quad (5)$$

Table 1
List of notations.

Notation	Description
n	Number of training examples
k	Code length
X_m	$n \times l_m$ feature matrix of modality m
A_m	$n \times n$ intra-modal graph of modality m
F_m	$n \times t_m$ topic matrix of modality m
F	$n \times t$ multi-modal topic matrix
H	$n \times k$ hash codes of training examples
P_m	$l_m \times k$ weight matrix of hash function for modality m
α_m	Weight of modality m in topic learning
β,θ_m	Regularization parameters

Substituting W_m in (4), the objective function becomes:

$$\sum_{m=1}^{M}\left(\mathrm{Tr}\left(F_m^T L_m F_m\right)+\alpha_m^2\,\mathrm{Tr}\left(I-F^T F_m F_m^T F\right)\right) \tag{6}$$

We adopt an alternating optimization to solve (6). More specifically, we alternatively update F, F_m and α_m to optimize the objective function.

(1) *Optimizing* F: We fix F_m and α_m, then (6) can be reformulated as:

$$\begin{aligned}&\max_{F}\quad \mathrm{Tr}\left(F^T\sum_{m=1}^{M}\left(\alpha_m^2 F_m F_m^T\right)F\right)\\ &\text{s.t.}\quad F^T F=I\end{aligned} \tag{7}$$

It is obviously that (7) is an eigenvalue problem, and we can obtain F by eigen-decomposition of $\sum_{m=1}^{M}\left(\alpha_m^2 F_m F_m^T\right)$.

(2) *Optimizing* F_m: We fix F and α_m. According to the trace property: $\mathrm{Tr}\left(F^T F_m F_m^T F\right)=\mathrm{Tr}\left(F_m^T F F^T F_m\right)$, (6) can be transformed to:

$$\begin{aligned}&\min_{F_m}\quad \mathrm{Tr}\left(F_m^T C_m F_m\right)\\ &\text{s.t.}\quad F_m^T F_m=I\end{aligned} \tag{8}$$

where

$$C_m=L_m-\alpha_m^2 F^T F \tag{9}$$

We can also find that F_m is learned by solving the eigenvalue problem of (8),

(3) *Optimizing* α_m: F and F_m are fixed, by using Lagrange multiplier, we can obtain:

$$\alpha_m=\frac{1/\mathrm{Tr}\left(I-F^T F_m F_m^T F\right)}{\sum_{i=1}^{M}1/\mathrm{Tr}\left(I-F^T F_i F_i^T F\right)} \tag{10}$$

The whole alternating optimization process is illustrated in Algorithm 1. In the implementation of this algorithm, we initialize F_m by solving the eigenvalue problem of (2), and $\alpha_m|_{m=1}^{M}$ are set to the same. Since the objective function is lower bounded by 0 and it will keep decreasing in each step, its convergence is guaranteed. One advantage of our topic learning is that the importance of different modality for generating the semantic topics is different, while previous cross-modal methods, such as CCA and IMH, treat all modalities equally. Thus our topic learning methods is more adaptive. Another advantage is that the hierarchical generation can effectively preserve both intra-modal and inter-modal consistency.

Algorithm 1. The learning process of HMMTL.

Input: $A_m|_{m=1}^{M}$
Output: F
1: Compute $L_m|_{m=1}^{M}$;
2: Initialize $F_m|_{m=1}^{M}$ and $\alpha_m|_{m=1}^{M}$;
3: **while** Not Converge **do**
4: Update F by solving the eigenvalue problem of (7);
5: Update $F_m|_{m=1}^{M}$ by solving the eigenvalue problem of (8);
6: Update $\alpha_m|_{m=1}^{M}$ according to (10);
7: **end while**

3.3. Robust matrix factorization for hash code generation

The eigenvalue decomposition in HMMTL obtains unbalanced topics. Generally, most semantic information is contained in top topics and the remaining topics usually contain less semantic information or even noises [8]. By choosing appropriate number of top multi-modal topics for F, we can effectively learn semantic information and remove noisy information. Then F can represent semantic correlation between different modalities.

However, due to the variance of semantic information in different topics, using each topic to generate one bit in hash codes is not reasonable [41,11]. Directly using F for hashing will result in much loss of semantic information in the quantization process. Therefore, we should design a hash generation process to effectively preserve the semantic information in quantized hash codes.

In this subsection, we introduce the Robust Matrix Factorization (RMF), which can effectively learn hash codes with balanced information. Matrix factorization has shown to be effective in the quantization of hash codes [39]. Unlike traditional matrix factorization, RMF uses $\ell_{2,1}$-norm to decompose F to H. According to the characteristic of $\ell_{2,1}$-norm [42,43], RMF is more robust to outliers, and it can better preserve the semantic correlation in hash codes H. The objective function of RMF for hash code learning is:

$$\min_{H,V}\|F-HV\|_{2,1}^2+\beta\left(\|H\|_F^2+\|V\|_F^2\right) \tag{11}$$

where $V\in\mathbb{R}^{k\times t}$, $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm, which is defined as $\|X\|_{2,1}=\sum_i\sqrt{\sum_j X_{ij}^2}$.

Then we can solve the $\ell_{2,1}$-norm problem. Eq. (11) can be transformed to:

$$\min_{H,V}\mathrm{Tr}\left((F-HV)^T D_H(F-HV)\right)+\beta\left(\|H\|_F^2+\|V\|_F^2\right) \tag{12}$$

where D_H is the diagonal matrix with its diagonal element $D_H^{ii}=1/2\left\|(F-HV)^i\right\|_2$

The above equation (12) can be also solved by the alternating process. In each iteration, we first optimize H and fix V, by setting the derivative of (12) w.r.t. H to zero, we can obtain:

$$HVV^T+\beta D_H^{-1}H=FV^T \tag{13}$$

Eq. (13) is the Sylvester equation, we can rewrite this

equation in the form:

$$\left(VV^T \otimes I + \beta I \otimes D_H^{-1}\right) vec(H) = vec\left(FV^T\right) \tag{14}$$

where $\otimes$ is the Kronecker product, and $vec(\cdot)$ is the vectorization operator. We can easily compute H by:

$$vec(H) = \left(VV^T \otimes I + \beta I \otimes D_H^{-1}\right)^{-1} vec\left(FV^T\right) \tag{15}$$

Then we fix H and optimize V, by setting the derivative of (12) w.r.t. V to zero, we can obtain V by:

$$V = \left(H^T D_H H + \beta I\right)^{-1} H^T D_H F \tag{16}$$

The alternative process of generating H is illustrated in Algorithm 2.

Algorithm 2. The learning process of RMF.

Input: F
Output: H
Initialize H and V randomly;
while Not Converge **do**
  Update H by (15);
  Update V by (16);
end while

3.4. Cross-modal hash function learning

In the previous steps we have learned the hash space which can effectively preserve the semantic correlation of multi-modal data. Then we formulate the hash function learning process which can project new examples into this hash space. One advantage of self-taught scheme is that different modalities are correlated in higher abstraction. Moreover, unlike previous cross-modal methods which have to jointly learn hash functions for all modalities, we can learn their hash functions separately. Once we have obtained the semantic hash space, then all modalities can be easily projected into this space which correlates them. We can even project new modalities into this space without any changes for the whole framework.

The hash projection process should be efficient, in that it occupies a certain part of the search time. For this purpose, we use the linear projection as hash function, which is defined as:

$$h_m = \mathrm{sgn}(x_m P_m - b_m) \tag{17}$$

where x_m is the feature vector of modality m. $P_m \in \mathbb{R}^{l_m \times k}$ is the hash projection matrix for modality m, k is the code length, and l_m is the feature dimension of modality m. b_m is the threshold parameter, it is computed as the mean of all $x_m P_m$ from the training data.

Our hash function learning is different to [35], in that we use soft assignment of H which is obtained from RMF. The hard assignment of H is suited for the online search process, since binary codes are more efficient. However, it will cause the loss of semantic information, thus it is not suited for learning process, which is done offline. As a result, our hash function learning becomes a regression problem, and we can use the following regularized least square regression for each modality:

$$\min_{P_m} \|X_m P_m - H\|_F^2 + \theta_m \|P_m\|_F^2 \tag{18}$$

From (18) we can easily obtain that:

$$P_m = \left(X_m^T X_m + \theta_m I\right)^{-1} X_m^T H \tag{19}$$

Algorithm 3. The overall learning process of CMSTH.

Input:
$A_m|_{m=1}^M$, $X_m|_{m=1}^M$
Output:
F, H, $P_m|_{m=1}^M$
Compute F by Algorithm 1;
Compute H by Algorithm 2;
for each modality m **do**
  Compute P_m by (19);
end for

The overall learning process of CMSTH is shown in Algorithm 3. Given a new example x_m, we directly use (17) to compute its hash vector h_m, which is then used to search any modalities by directly deploying the hamming distance.

4. Experiments

4.1. Datasets and features

In this paper, two real world multi-modal image datasets: Wikipedia [3] and NUS-WIDE [44] are used for evaluation. These two datasets are both split into independent training set and test set. Training set is used to learn hash functions of all methods, and the retrieval performance is evaluated on test set. The statistics of two datasets are summarized in Table 2.

Wikipedia dataset was assembled from the "Wikipedia feature articles". It contains 2866 multi-modal documents (image-text pairs), and each of them is labeled with exactly one of 10 semantic labels. All labels are only used as ground truth, they are not used for training. Documents which share the same concept are regarded as relevant. 2173 image-text pairs in Wikipedia dataset are chosen as training set, and the rest 693 pairs are used as test set.

NUS-WIDE dataset contains 269,648 multi-modal documents, each multi-modal document is also an image-text pair and text in NUS-WIDE refers to the associated social tags. The image-text pairs are labeled by 81 concepts that are only used for evaluation. We prune the original NUS-WIDE to form a new dataset consisting of 203,597 image-text pairs by keeping the images that have at least one tag and one concept. Then this dataset is split into 5090 training set and 198,507 test set.

On Wikipedia, we extract 1000-D SIFT histogram, 1000-D HOG histogram, 500-D GIST for image, and extract 6000-D tf-idf vector for text. On NUS-WIDE, we directly use six image features and one binary text feature provided by [44].[1]

Since some compared methods are not suited to multiple image features, we use Kernel PCA (KPCA) [45] to

[1] All features can be downloaded from http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Table 2
The statistics of two datasets.

Datasets	Wikipedia	NUS-WIDE
Training examples	2173	5090
Test images	693	198,507
Test texts	693	198,507
Dimension of image feature	200	100
Dimension of text feature	30	1000

combine image features and reduce their dimensions. Histogram intersection kernel is used for all image histograms and BoVWs, and RBF Kernel is used for other visual features. All visual kernels are linearly combined to train the KPCA. Finally we obtain 200-D visual feature for images on Wikipedia, and 100-D visual feature for images on NUS-WIDE. Since the dimension of text feature on Wikipedia is too high, we also use KPCA to reduce its dimension to 30, and histogram intersection kernel is used.

4.2. Evaluation metrics

We adopt non-interpolated Mean Average Precision (MAP) to measure the retrieval performance. Given a query and a list of R retrieved results, the Average Precision (AP) is defined as:

$$AP = \frac{1}{p}\sum_{i=1}^{R} pre(i)rel(i) \tag{20}$$

where p is the number of relevant documents in the retrieved set, $pre(i)$ is the precision of top i retrieved documents. $rel(i)=1$ if the ith retrieved documents is relevant to query, otherwise $rel(i)=0$. The MAP score is the mean of AP scores from all the queries. In our work, we set $R=50$, thus the MAP scores are computed on the top 50 retrieved documents of each query. Besides MAP, we use Precision–Recall (PR) curves to measure the retrieval performance.

4.3. Compared methods and implementation details

We compare our method with four representative unsupervised cross-modal hashing methods, including CVH [9],[2] IMH [40], CHMIS [10], CMFH [39], and two unimodal image hashing methods, including SH [33] and STH [35]. The codes of all compared methods are publicly available. For all methods, we choose the parameters which make them perform best.

We also introduce a Baseline which is used to show the advantage of RMF in our CMSTH. In Baseline, the first and third steps of CMSTH are preserved, and RMF is removed. Therefore, the baseline directly uses multi-modal topics as hash codes.

In the implementation of CMSTH, we choose the parameters that make CMSTH perform best or nearly best. On both two datasets, the nearest neighbors c for $A_m|_{m=1}^{M}$ are set to 500, and we set all $\theta_m = 1, \beta = 0.1$. The number of unimodal and multi-modal topics is the same, they are set to 8 on Wikipedia dataset, and 30 on NUS-WIDE.

[2] We use the codes implemented by http://www.cse.ust.hk/~dyyeung/code/mlbe.zip

Table 3
MAP scores of cross-modal retrieval on Wikipedia.

Wiki	Method	Code length			
		16	32	64	128
Image query	CVH	0.2382	0.2377	0.2080	0.1920
	IMH	0.2554	0.2694	0.2615	0.2454
	CHMIS	0.2346	0.2365	0.2194	0.1819
	CMFH	0.2857	0.2995	0.3041	0.3019
	Baseline	0.2811	0.2797	0.2522	0.2298
	CMSTH	**0.3155**	**0.3293**	**0.3313**	**0.3375**
Text query	CVH	0.2644	0.2606	0.2274	0.2003
	IMH	0.2799	0.2905	0.2806	0.2634
	CHMIS	0.2781	0.2572	0.2341	0.1922
	CMFH	0.3087	0.3242	0.3296	0.3330
	Baseline	0.2976	0.2932	0.2684	0.2395
	CMSTH	**0.3562**	**0.3700**	**0.3825**	**0.3878**

4.4. Results of cross-modal image retrieval

In the cross-modal image retrieval, all methods only use training data to learn hash functions, which are then used to obtain hash codes for test data. We evaluate two types of cross-modal retrieval tasks, one is Image Query, where test images are used to search test texts; the other is Text Query, where test texts are used to search test images.

Table 3 shows the MAP scores of all cross-modal methods on Wikipedia dataset. In Image Query, all test images are used as queries, and test texts form database. In Text Query, all test texts are chosen as queries and test images form the database. From Table 3 we can find that CMSTH performs best in all cases, which confirms the advantages of the analyzing semantic correlation. When the code length is 16, CMSTH obtains relatively high MAP scores. If retrieval time is the major concern, CMSTH can guarantee the performance by adopting small code length. Moreover, CMSTH consistently improves the retrieval performance by increasing the code length.

We also find that CMSTH significantly outperforms Baseline which directly uses multi-modal topics as hash codes and do not use RMF. The results confirm that RMF is effective in generating hash codes. Although multi-modal topics can well represent the semantic correlation of different modalities, they are not suited for quantization. RMF can losslessly preserve the semantic information of multi-modal topics in final codes, thus it significantly improves the hashing performance. CVH and CHMIS do not consider the semantic correlation in multi-modal data, thus they perform worse than Baseline and CMSTH. CMFH and IMH also ignore the semantic correlation, but they can generate effective codes for quantization, so they are slightly better than Baseline in some cases. The PR curves of all cross-modal methods on Wikipedia are shown in Fig. 2, we can find the results are consistent with MAP scores.

Table 4 shows the MAP scores of all cross-modal methods on NUS-WIDE dataset. In Image Query, we choose 1% of test images as queries, and all test texts form

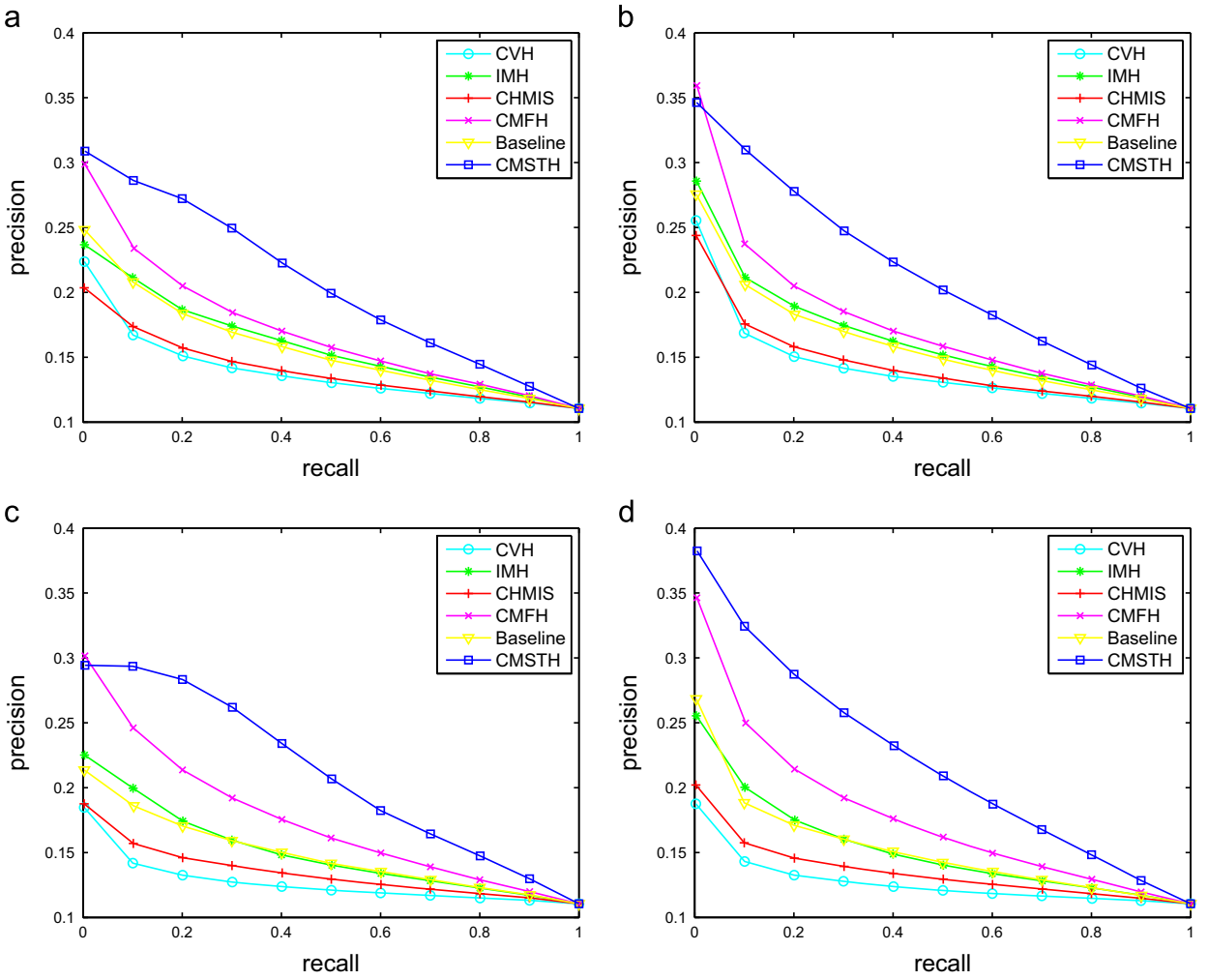

Fig. 2. PR curves of cross-modal retrieval on Wikipedia: (a) Image query $k=32$; (b) text query $k=32$; (c) image query $k=64$; and (d) text query $k=64$.

Table 4
MAP scores of cross-modal retrieval on NUS-WIDE.

Wiki	Method	Code length			
		16	32	64	128
Image query	CVH	0.4441	0.4371	0.4229	0.3966
	IMH	0.3401	0.3507	0.3637	0.3808
	CHMIS	0.4155	0.4133	0.3952	0.3597
	CMFH	0.3303	0.3894	0.3837	0.3981
	Baseline	0.4408	0.4588	0.5048	0.5031
	CMSTH(3)	**0.5032**	**0.5073**	**0.5270**	**0.5439**
Text query	CVH	0.4065	0.4136	0.4110	0.3846
	IMH	0.3337	0.3557	0.3699	0.3764
	CHMIS	0.3972	03966	0.3899	0.3562
	CMFH	0.4004	0.4078	0.4174	0.4192
	Baseline	0.4447	0.4486	0.4918	0.4906
	CMSTH(3)	**0.4761**	**0.4965**	**0.5088**	**0.5243**

database. In Text Query, 1% of test texts is chosen as queries and all test images form the database. The results on NUS-WIDE are similar to Wikipedia, CMSTH performs best in all cases, which further confirms the advantages of our method. CMSTH also performs better than Baseline, which illustrates that RMF can improve the performance of hashing. The only difference is that Baseline performs better than other methods except for CMSTH. The reason may be that semantic correlation is more important for NUS-WIDE, and both Baseline and CMSTH can well analyze the semantic correlation. The PR curves of all cross-modal methods on NUS-WIDE are shown in Fig. 3, and the results are also consistent with MAP scores.

4.5. *Results of unimodal image retrieval*

We then show the results of cross-modal methods on unimodal image retrieval. Unimodal retrieval is also practical in some applications, using image queries to search images is required by users sometimes. Therefore, we evaluate our method in unimodal image retrieval. In each retrieval process, we choose one test image as a query, and other test images form database. All test images are chosen as queries on Wikipedia, and 1% of test images is chosen as queries on NUS-WIDE. Generally, cross-modal methods should perform better than unimodal methods in

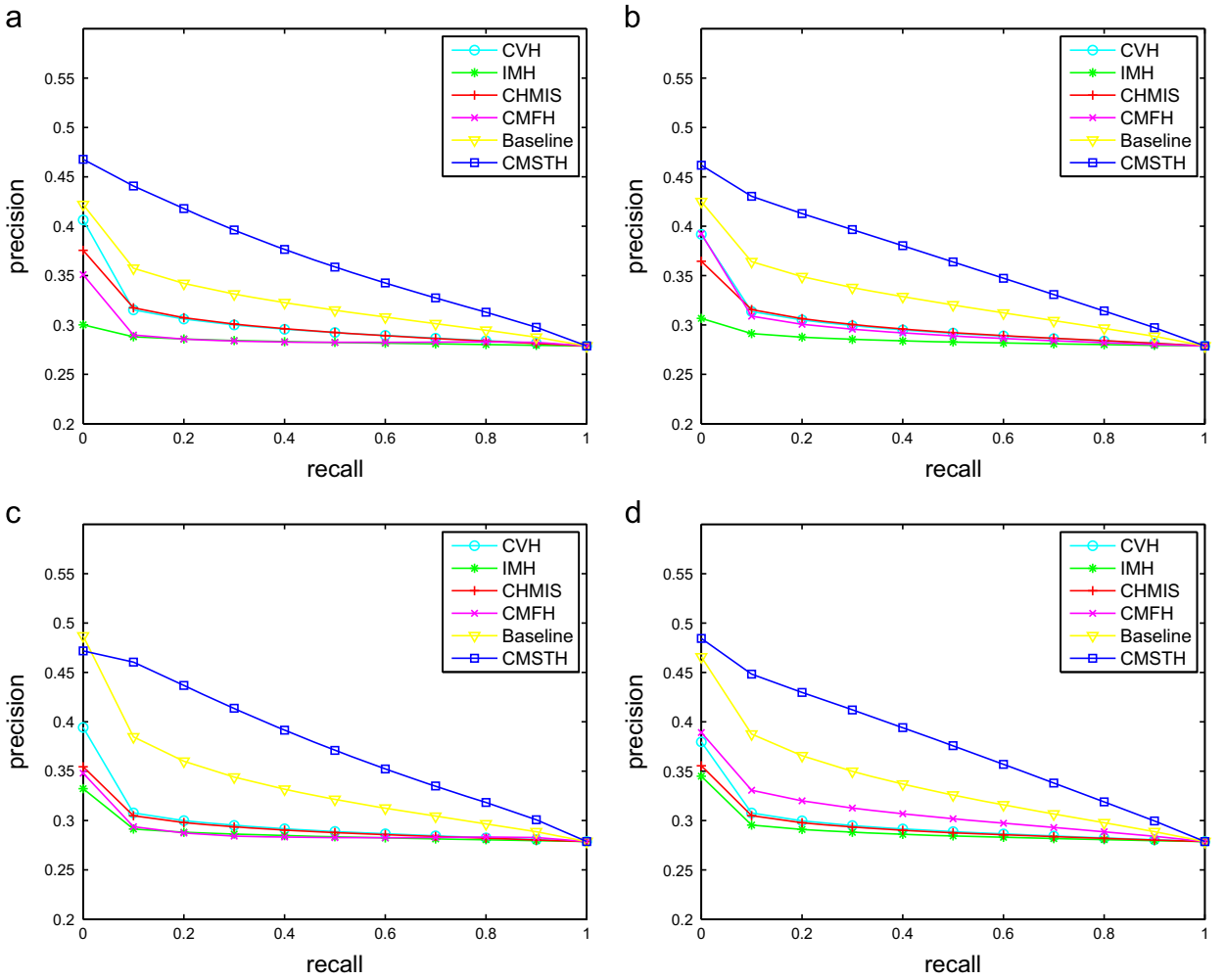

Fig. 3. The PR curves of cross-modal retrieval on NUS-WIDE: (a) Image query $k=32$; (b) text query $k=32$; (c) image query $k=64$; and (d) text query $k=64$.

Table 5
MAP scores of unimodal image retrieval on two datasets.

Method	Wikipedia				NUS-WIDE			
	16	32	64	128	16	32	64	128
SH	0.3734	0.3779	0.3879	0.4058	0.4671	0.5855	0.5994	0.5999
STH	0.4004	0.4039	0.4083	0.4047	0.4660	0.4813	0.5002	0.6228
CVH	0.3879	0.3862	0.3877	0.3839	0.5522	0.6201	0.6326	0.6528
IMH	0.3912	0.4046	0.3970	0.4011	0.4461	0.5447	0.5894	0.6421
CHMIS	0.3920	0.3934	0.3999	0.4089	0.5177	0.6116	0.6414	0.6541
CMFH	0.3988	0.3978	0.4046	0.4150	0.3718	0.3789	0.4729	0.5764
CMSTH	**0.4090**	**0.4326**	**0.4344**	**0.4492**	**0.5666**	**0.6584**	**0.6731**	**0.6876**

image retrieval, in that exploiting associated text can better understand the semantics of images.

Table 5 shows the results of all compared hashing methods on two datasets, we can find CMSTH that obtains the highest MAP scores in all cases, which shows that CMSTH improves the performance of unimodal image retrieval. However, the advantage of other cross-modal methods, including CVH, IMH, CHMIS and CMFH is not significant. They even perform worse than unimodal hashing methods SH and STH. This phenomenon illustrates that cross-modal methods may not always improve the performance of unimodal retrieval. Both CMSTH and baseline analyze semantic correlation, their better results confirm that semantic correlation of multi-modal data also

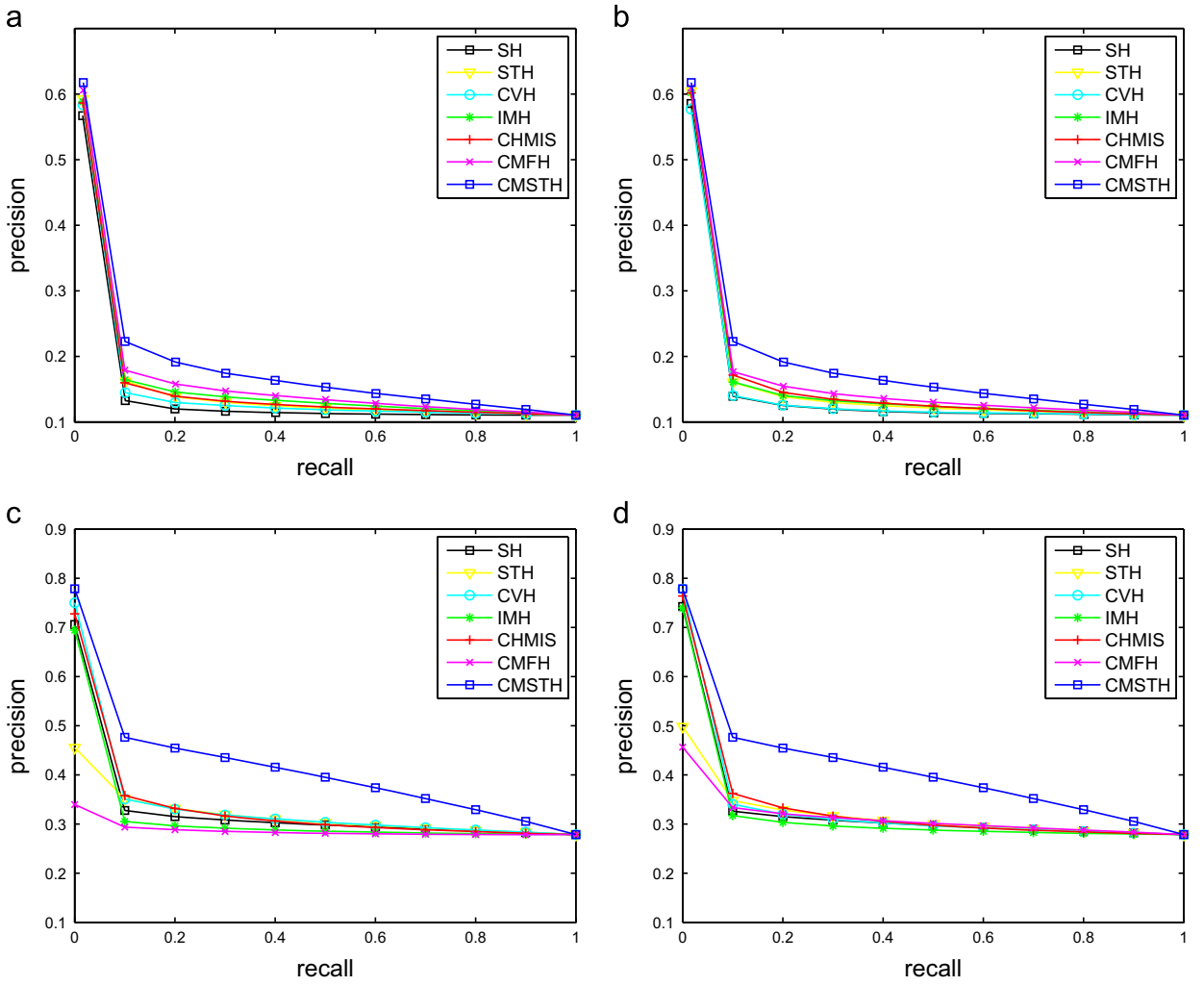

Fig. 4. The PR curves of unimodal image retrieval on two datasets: (a) Wikipedia $k=32$; (b) Wikipedia $k=64$; (c) NUS-WIDE $k=32$; and (d) NUS-WIDE $k=64$.

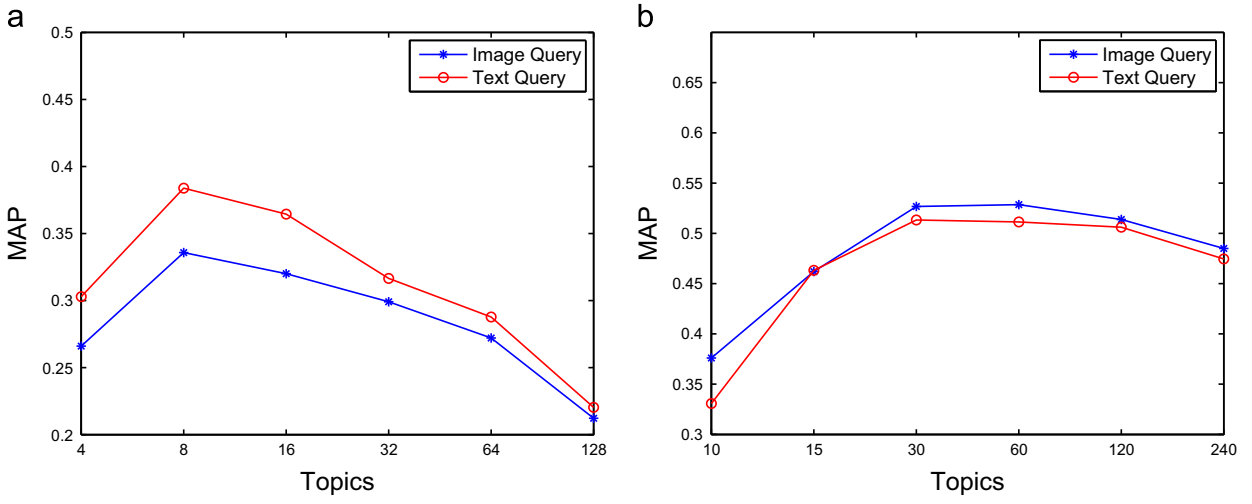

Fig. 5. Performance variations with different number topics in CMSTH on two datasets, the code length is 64: (a) Wikipedia and (b) NUS-WIDE.

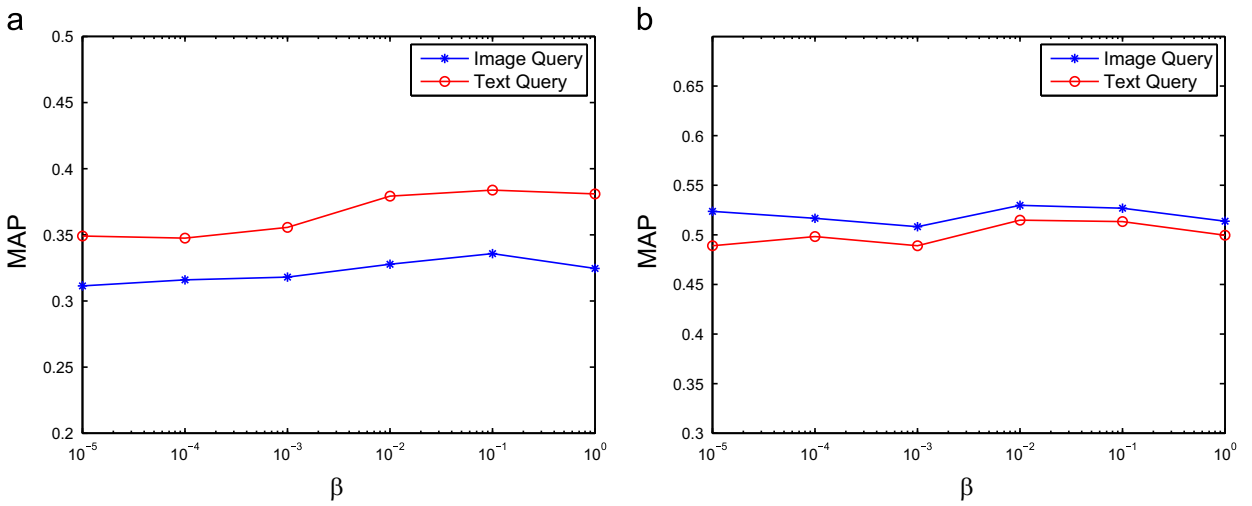

Fig. 6. Performance variations with different value of β in CMSTH on two datasets, the code length is 64: (a) Wikipedia and (b) NUS-WIDE.

well represents the relation of unimodal data. The PR curves of unimodal retrieval are shown in Fig. 4, and the results are consistent with MAP scores.

4.6. Parameter analysis

At last we analyze the influence of parameters in CMSTH. The nearest neighbors c and regularization parameters θ_m have been comprehensively discussed by previous graph learning and regularized least square methods [10,31,43], thus we do not focus on their analysis.

We mainly analyze the influence of topic number and β which are more important in CMSTH. In our experiment, we set the same number for unimodal and multi-modal topics. Fig. 5 shows the MAP score variations with a different number of topics on two datasets, and the code length k is set to 64. We set topic number as $\{4, 8, 16, 32, 64, 128\}$ on Wikipedia, and $\{10, 15, 30, 60, 120, 240\}$ on NUS-WIDE. From this figure we can observe that CMSTH is sensitive to the topic numbers, it obtains highest scores with topic number 8 on Wikipedia, and 30 on NUS-WIDE. The results also illustrate that larger number of topics does not mean the better performance of cross-modal retrieval. They confirm that the semantic information can be effectively learned and noises can be removed by eigenvalue decomposition.

Fig. 6 shows MAP score variations with a different value of β on two datasets, and k is set to 64. From the figure we can find that our method is not very sensitive to β. The highest MAP scores are obtained at $\beta=0.1$ on Wikipedia and $\beta=0.01$ on NUS-WIDE, respectively. For $\beta=0.1$ on NUS-WIDE, the MAP score is also relatively high, thus in our experiments we set $\beta=0.1$ on both two datasets.

5. Conclusions and future work

In this paper we introduce Cross-Modal Self-Taught Hashing (CMSTH) for both cross-modal and unimodal image retrieval. CMSTH can correlate different modalities by semantic topics, while previous unsupervised methods only analyze the content correlation. In the first step of CMSTH, we propose Hierarchical Multi-Modal Topic Learning (HMMTL) to learn topics of multi-modal data. In the learning process of topics, both intra-modal and inter-modal consistency are preserved, and proper weights are allocated for different modalities. Then we use Robust Matrix Factorization (RMF) to transfer the topics to hash codes which are more suited to quantization. In the last step, we learn the hash functions to make all modalities be correlated in the obtained hash space. Experimental results on Wikipedia and NUS-WIDE show that CMSTH significantly outperforms other cross-modal and unimodal hashing methods.

A potential advantage of CMSTH is that it can be easily extended without any changes in the whole framework. For example, we may introduce a new topic model for first step, to learn topics with more semantic information. In the last step, most existing state-of-the-art supervised methods can be used to learn the hash functions. In the future work, we will consider improving the topic learning and hash function learning steps.

References

[1] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. (CSUR) 40 (2) (2008) 5.

[2] Y. Yang, D. Xu, F. Nie, J. Luo, Y. Zhuang, Ranking with local regression and global alignment for cross media retrieval, in: ACM Multimedia, ACM, Beijing, China, 2009, pp. 175–184.

[3] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: ACM Multimedia, ACM, Firenze, Italy, 2010, pp. 251–260.

[4] L. Xie, P. Pan, Y. Lu, A semantic model for cross-modal and multi-modal retrieval, in: ACM ICMR, ACM, Dallas, USA, 2013, pp. 175–182.

[5] S.J. Hwang, K. Grauman, Learning the relative importance of objects from tagged images for retrieval and cross-modal search, Int. J. Comput. Vis. 100 (2) (2012) 134–153.

[6] K. He, F. Wen, J. Sun, K-means hashing: an affinity-preserving quantization method for learning binary compact codes, in: CVPR, IEEE, 2013, pp. 2938–2945.

[7] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: ICCV, IEEE, 2009, pp. 2130–2137.

[8] J. Cheng, C. Leng, P. Li, M. Wang, H. Lu, Semi-supervised multi-graph hashing for scalable similarity search, Comput. Vis. Image Underst. 124 (2014) 12–21.
[9] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: IJCAI, vol. 22, 2011, p. 1360.
[10] D. Zhang, F. Wang, L. Si, Composite hashing with multiple information sources, in: ACM SIGIR, ACM, Beijing, China, 2011, pp. 225–234.
[11] D. Zhang, W.-J. Li, Large-scale supervised multimodal hashing with semantic correlation maximization, in: AAAI, 2014, pp. 2177–2183.
[12] 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.
[13] D. Li, N. Dimitrova, M. Li, I.K. Sethi, Multimedia content processing through cross-modal association, in: ACM Multimedia, ACM, Berkeley, USA, 2003, pp. 604–611.
[14] L. Xie, P. Pan, Y. Lu, S. Jiang, Cross-modal self-taught learning for image retrieval, in: MultiMedia Modeling, Springer, 2015, pp. 257–268.
[15] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, arXiv preprint arXiv:1304.5634.
[16] Y. Luo, D. Tao, B. Geng, C. Xu, S.J. Maybank, Manifold regularized multitask learning for semi-supervised multilabel image classification, IEEE Trans. Image Process. 22 (2) (2013) 523–536.
[17] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, Y. Wen, Multiview vector-valued manifold regularization for multilabel image classification, IEEE Trans. Neural Netw. Learn. Syst. 24 (5) (2013) 709–722.
[18] W. Liu, D. Tao, Multiview Hessian regularization for image annotation, IEEE Trans. Image Process. 22 (7) (2013) 2676–2687.
[19] Y. Luo, T. Liu, D. Tao, C. Xu, Multiview matrix completion for multilabel image classification, IEEE Trans. Image Process. 24 (8) (2015) 2355–2368.
[20] Y. Luo, T. Liu, D. Tao, C. Xu, Decomposition-based transfer distance metric learning for image classification, IEEE Trans. Image Process. 23 (9) (2014) 3789–3801.
[21] L. Xie, P. Pan, Y. Lu, S. Wang, A cross-modal multi-task learning framework for image annotation, in: ACM CIKM, ACM, Shanghai, China, 2014, pp. 431–440.
[22] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, IEEE Trans. Pattern Anal. Mach. Intell. (2015) 1.
[23] H. Hotelling, Relations between two sets of variates, Biometrika (1936) 321–377.
[24] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. 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.
[25] Y. Yang, Y.-T. Zhuang, F. Wu, Y.-H. Pan, Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval, IEEE Trans. Multimed. 10 (3) (2008) 437–446.
[26] V. Mahadevan, C.W. Wong, J.C. Pereira, T. Liu, N. Vasconcelos, L.K. Saul, Maximum covariance unfolding: manifold learning for bimodal data, in: NIPS, 2011, pp. 918–926.
[27] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: NIPS, 2012, pp. 2222–2230.
[28] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: ACM Multimedia, ACM, Orlando, USA, 2014, pp. 7–16.
[29] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, Y. Zhuang, A low rank structural large margin method for cross-modal ranking, in: ACM SIGIR, ACM, Dublin, Ireland, 2013, pp. 433–442.
[30] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, Y. Zhuang, Cross-media semantic representation via bi-directional learning to rank, in: ACM Multimedia, ACM, Barcelona, Spain, 2013, pp. 877–886.
[31] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: ACM SIGMOD, ACM, New York, USA, 2013, pp. 785–796.
[32] A. Gionis, P. Indyk, R. Motwani, et al., Similarity search in high dimensions via hashing, in: VLDB, vol. 99, 1999, pp. 518–529.
[33] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: NIPS, 2009, pp. 1753–1760.
[34] W. Liu, J. Wang, S. Kumar, S.-F. Chang, Hashing with graphs, in: ICML, 2011, pp. 1–8.
[35] D. Zhang, J. Wang, D. Cai, J. Lu, Self-taught hashing for fast similarity search, in: ACM SIGIR, ACM, Geneva, Switzerland, 2010, pp. 18–25.
[36] M.M. Bronstein, A.M. Bronstein, F. Michel, N. Paragios, Data fusion through cross-modality metric learning using similarity-sensitive hashing, in: CVPR, IEEE, 2010, pp. 3594–3601.
[37] Y. Zhen, D.-Y. Yeung, A probabilistic model for multimodal hash function learning, in: ACM SIGKDD, ACM, Beijing, China, 2012, pp. 940–948.
[38] X. Zhu, Z. Huang, H.T. Shen, X. Zhao, Linear cross-modal hashing for efficient multimedia search, in: ACM Multimedia, ACM, Barcelona, Spain, 2013, pp. 143–152.
[39] G. Ding, Y. Guo, J. Zhou, Collective matrix factorization hashing for multimodal data, in: CVPR, IEEE, 2014, pp. 2083–2090.
[40] X. Zhai, Y. Peng, J. Xiao, Cross-media retrieval by intra-media and inter-media correlation mining, Multimed. Syst. 19 (5) (2013) 395–406.
[41] J. Wang, S. Kumar, S.-F. Chang, Semi-supervised hashing for large-scale search, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2393–2406.
[42] Z. Ma, Y. Yang, N. Sebe, A.G. Hauptmann, Knowledge adaptation with partially shared features for event detection using few exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 36 (9) (2014) 1789–1802.
[43] J. Song, Y. Yang, X. Li, Z. Huang, Y. Yang, Robust hashing with local models for approximate similarity search, IEEE Trans. Cybern. 44 (7) (2014) 1225–1236.
[44] T.-S. 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: ACM CIVR, ACM, Santorini Island, Greece, 2009, p. 48.
[45] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.