Semantic Topic Multimodal Hashing for Cross-Media Retrieval

Di Wang, Xinbo Gao, Xiumei Wang, Lihuo He
School of Electronic Engineering, Xidian University
Xi'an, 710071, China
wangdi.wandy@gmail.com, xbgao@mail.xidian.edu.cn,
wangxm@xidian.edu.cn, lihuo.he@gmail.com

Abstract

Multimodal hashing is essential to cross-media similarity search for its low storage cost and fast query speed. Most existing multimodal hashing methods embedded heterogeneous data into a common low-dimensional Hamming space, and then rounded the continuous embeddings to obtain the binary codes. Yet they usually neglect the inherent discrete nature of hashing for relaxing the discrete constraints, which will cause degraded retrieval performance especially for long codes. For this purpose, a novel Semantic Topic Multimodal Hashing (STMH) is developed by considering latent semantic information in coding procedure. It first discovers clustering patterns of texts and robust factorizes the matrix of images to obtain multiple semantic topics of texts and concepts of images. Then the learned multimodal semantic features are transformed into a common subspace by their correlations. Finally, each bit of unified hash code can be generated directly by figuring out whether a topic or concept is contained in a text or an image. Therefore, the obtained model by STMH is more suitable for hashing scheme as it directly learns discrete hash codes in the coding process. Experimental results demonstrate that the proposed method outperforms several state-of-the-art methods.

1 Introduction

With the rapid development of the Internet and multimedia devices such as smart phone, tablet computer, and digital camera, tremendous amounts of multimedia data including texts, images, and videos have been easily obtained. It is common that relevant multimedia data from different media types may have semantic correlations. For example, a microblog in Facebook often consists of a short text and some correlative images; a video in YouTube is always associated with some related descriptions or tags. As semantic information inherently consists of data with different modalities, it gives rise to an emerging demand to study and explore the interactions between multimodal data for the applications like cross-media retrieval [Zhai *et al.*, 2013], image annotation [Weston *et al.*, 2011], and recommendation system [Bedi *et al.*, 2007]. Multimodal hashing methods, which map heterogeneous data points into a common low-dimensional Hamming space, have recently received considerable attentions to address cross-modality similarity search problem. The core problem of multimodal hash codes learning is how to construct the underlying correlations between the multiple modalities and preserve the similarity relationships in each individual modality. Generally, multimodal hashing methods can be divided into two categories: graph based methods and matrix decomposition based ones.

Graph based multimodal hashing methods construct similarity graph for each individual modality to preserve the intra-modal similarities and simultaneous concatenate multiple modality-specific binary codes to preserve the inter-modal similarities for the final hash codes. The learning problems often can be converted into eigen-decomposition problems by means of relaxation. Cross-view hashing (CVH) [Kumar and Udupa, 2011] extends spectral hashing (SH) [Weiss *et al.*, 2009] to the multimodal setting by maximizing the weighted average correlations between data pairs through solving a generalized eigenvalue problem. However, CVH treats the correlations between inter-classes and intra-classes in the same way and such strategy often results in poor performance as the differences between the modalities are ignored. Inter-media hashing (IMH) [Song *et al.*, 2013] takes the differences between multiple modalities into consideration. It first explores the correlations within each single modality according to similarity graph and then keeps the binary codes of the paired data points with different modalities consistent. However, IMH needs to construct the similarity matrix for all the data points, which will lead to a large computational complexity for large-scale data set. Linear cross-modal hashing (LCMH) [Zhu *et al.*, 2013] avoids the large-scale graph construction by representing training data with a small number of cluster centers. Nevertheless, how to choose appropriate cluster centers of massive data set is a difficult problem as the performance of LCMH is sensitive to the number of clusters. Generally, those graph based multimodal hashing methods have two drawbacks. Firstly, considerable computational complexity for computing similarity graph leads to long training time. Secondly, eigen-decomposition process decreases the mapping quality substantially when increasing the number of bits, since most of the information is contained in the top few eigenvectors.

Matrix decomposition based multimodal hashing methods seek latent low dimensional spaces to well reconstruct multimodal data and quantify the reconstruction coefficients to obtain the binary codes. Such kinds of methods can avoid the large scale graph construction and eigen-decomposition. Collective matrix factorization hashing (CMFH) [Ding *et al.*, 2014] learns the unified hash codes by collective matrix factorization with latent factor model from different modalities of one instance. However, it assumes that each view of one instance generates identical hash codes. This identical constraint is too restrictive which only guarantees the consistence of pairwise data points but ignores the cross correlationship between different pairwise data points. Latent semantic sparse hashing (LSSH) [Zhou *et al.*, 2014] uses sparse coding and matrix factorization to learn the latent spaces and then merges the learned latent features to generate binary codes. As sparse coding results in long training time consuming, LSSH cannot fully utilize all the training data to learn the model, thus will deteriorate the accuracy.

Although the above multimodal hashing methods have achieved promising results in multimodal applications, these methods often discard the discrete constraints of hashing to solve a relaxed problem and afterwards round the continuous solutions to get the binary codes. Whereas, such continuous relaxation scheme will cause large quantization error consequently deteriorate search performance especially for long codes length. To this end, we design a novel multimodal hashing model, namely semantic topic multimodal hashing (STMH), which focuses on the binary nature of hashing, to facilitate the large-scale cross-media retrieval for multimedia data sources with texts and images. As illustrated in Figure 1, STMH represents text in (a) as multiple semantic topics in (c) and image in (b) as multiple semantic concepts in (d). After represented both texts and images in latent semantic spaces, STMH transforms the learned multimodal semantic features to a common subspace by their semantic correlations. Finally, it generates the unified hash codes directly by figuring out whether a topic or concept is contained in a text or an image. The contributions of the proposed model are summarized as below:

- The proposed method explicitly considers the implication and preserves the discrete nature of the hash code. Each bit of hash code represents whether a text or an image contains the corresponding topic or concept. By focusing on the binary nature, STMH is more suitable for hashing learning scheme and achieves better search performance.
- We develop a fast and efficient method which learns latent topics from texts and the unified hash codes simultaneously.
- By using $\ell_{2,1}$-norm to learn the robust salient concepts of images, the proposed method can achieve better robustness against noisy and unreliable image data.

The rest of this paper is organized as follows. Section 2 introduces the proposed STMH model and its multiple modalities extension. Experimental results and comparisons on two benchmark data sets are presented in Section 3. Finally, the conclusions are given in Section 4.

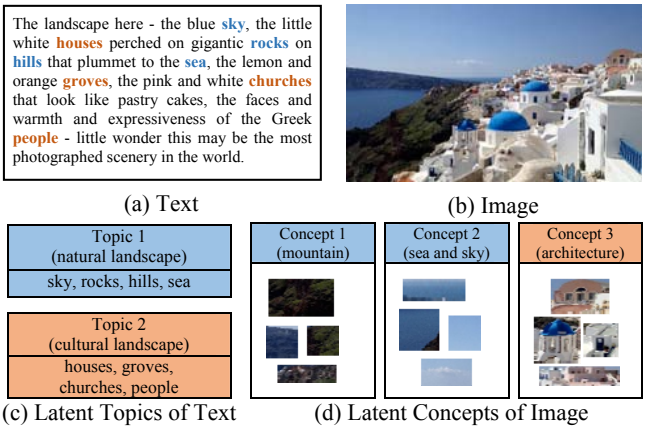

Figure 1: An illustration of topics and concepts for text and image about scenery of Santorini from Wikipedia.

2 Semantic Topic Multimodal Hashing

This section details the proposed STMH model. We firstly introduce STMH to bimodal instance consisting of images and texts as they are the most common-used and important modalities in multimedia. Without loss of generality, it can be easily extended to cases with more modalities.

2.1 Problem Formulation

Suppose that $\mathcal{O}=\{o_i\}_{i=1}^n$, $o_i=(\mathbf{x}_i,\mathbf{y}_i)$ is a set of multimodal objects, where $\mathbf{x}_i \in \mathbb{R}^m$ is a m-dimensional image feature, and $\mathbf{y}_i \in \mathbb{R}^d$ is a d-dimensional text feature (usually, $m \neq d$). Given the bits length k, the purpose of STMH is to learn an integrated binary code $\mathbf{h}_i \in \{0,1\}^k$ for o_i, $i=1,2,\cdots,n$, such that $\mathbf{h}_i$ and $\mathbf{h}_j$ preserve the semantic similarity between o_i and o_j with high probability. More specifically, if o_i and o_j are two objects have similar semantic, $\mathbf{h}_i$ and $\mathbf{h}_j$ should have a small Hamming distance, and vice versa.

2.2 Semantic Modeling

Semantic information usually can be extracted from a large data set by the topic model [Blei, 2012]. The basic idea of the topic model [Blei *et al.*, 2003; Blei, 2012] is that data are represented as random mixtures over an underlying set of topics, where each topic is characterized by a distribution over words. There are at least two advantages of modeling data as a set of topics for retrieval. Firstly, it offers a high-level abstraction which can remove redundant information or noise and highlight the important information for complex data. Secondly, the data can be described by a small number of semantic topics which make the computation of the semantic similarity between a data-pair fast and accurately. Motivated by topic models, STMH discovers latent semantic topics in texts and explores latent semantic concepts in images.

Semantic Topic Modeling for Text

To extract the semantic topics, the paper discovers latent topic patterns via a way like cluster analysis. And texts are projected to the latent topic space formed by cluster centers.

Then a text can be described as the topic distribution which can be easily represented by hash codes. When a topic is contained in a text, the corresponding hash code is 1, and otherwise 0.

Let $\mathbf{Y}$ be a set of d-dimensional text features, i.e., $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n] \in \mathbb{R}^{d\times n}$, $\mathbf{H}$ be the corresponding unified hash codes for $\mathcal{O}$, i.e., $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n] \in \mathbb{R}^{k\times n}$, and $\mathbf{F}$ be a set of d-dimensional latent semantic topics, i.e., $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_k] \in \mathbb{R}^{d\times k}$. The semantic topic modeling for text is

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{H}} \sum_{j=1}^{n}\sum_{i=1}^{k} h_{ij} \left\|\mathbf{y}_j - \mathbf{f}_i\right\|_2^2 \\ &s.t.\ h_{ij} \in \{0,1\}, \sum_{i=1}^{k} h_{ij} = c\ , \end{aligned} \tag{1}$$

where h_{ij} is the i-th element of $\mathbf{h}_j$ and c is a positive integer less than k which denotes how many "1" lies in $\mathbf{h}_j$. In this paper, c is set to $k/2$ to balance partition binary codes in order to maximize the information of each bit. Eq. (1) is used to simultaneously learn the latent semantic topics and determine the closely related topics for a text adaptively.

Note that when each $\mathbf{f}_i$ in Eq. (1) has equal value, which means all the topics are the same, then the corresponding hash codes will randomly distribute. Therefore, the semantic similarities between texts will cannot be reflected by hash codes. To avoid this case, we consider to add the diversity regularization term of topics. Let $\mathbf{h}^i$ denotes the i-th row of hash codes $\mathbf{H}$, then it represents the distribution of the i-th topic $\mathbf{f}_i$ on the text set $\mathbf{Y}$. If $\mathbf{h}^i$ is in close proximity to $\mathbf{h}^j$, $i, j \in 1, 2, \cdots, k$, $i \neq j$, the value of $\mathbf{h}^i\left(\mathbf{h}^j\right)^T$ will be large, and $\mathbf{f}_i$ and $\mathbf{f}_j$ will be similar to each other with high probability. To avoid $\mathbf{f}_i$ and $\mathbf{f}_j$ equal to each other, we add the diversity regularization term of topics as follows.

$$\min_{\mathbf{F}} \sum_{i,j=1}^{k} w_{ij} \mathbf{f}_i{}^T \mathbf{f}_j = Tr\left(\mathbf{F}\mathbf{W}\mathbf{F}^T\right), \tag{2}$$

where

$$w_{ij} = \begin{cases} \mathbf{h}^i\left(\mathbf{h}^j\right)^T, & i \neq j \\ 0, & i = j. \end{cases} \tag{3}$$

$\mathbf{f}_i{}^T\mathbf{f}_j$ is the dot-product similarity metric which has been widely used in text analysis [Cai *et al.*, 2011]. Therefore, it is quite natural to use dot-product similarity metric for topics. When w_{ij} has a relatively large value, it means that the distributions of topics $\mathbf{f}_i$ and $\mathbf{f}_j$ are similar to each other. To minimize Eq. (2), $\mathbf{f}_i{}^T\mathbf{f}_j$ should be small, then topics $\mathbf{f}_i$ and $\mathbf{f}_j$ should be different from each other. Therefore, the diversity of topics is increased.

Accordingly, the overall objective function for text semantic modeling can be obtained by combining Eqs. (1) and (2)

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{H}} \mathcal{L}_T = \sum_{j=1}^{n}\sum_{i=1}^{k} h_{ij} \left\|\mathbf{y}_j - \mathbf{f}_i\right\|_2^2 + Tr\left(\mathbf{F}\mathbf{W}\mathbf{F}^T\right) \\ &s.t.\ h_{ij} \in \{0,1\}, \sum_{i=1}^{k} h_{ij} = c\ . \end{aligned} \tag{4}$$

By minimizing Eq. (4), the latent semantic topics for text and the corresponding hash codes will be learned simultaneously.

Semantic Concept Modeling for Image

Compared with text, the high-level semantic concepts hidden in the image are more difficult to extract. Here, we use matrix factorization to discover semantic concepts in image. Matrix factorization which learns a latent low dimensional space to well reconstruct the original data, is one of the most useful tools for learning latent concepts from image [Lee and Seung, 1999; Zhou and Tao, 2011]. However, the standard matrix factorization often uses the least square error function which is well known to be unstable w.r.t. noise and outliers [Kong *et al.*, 2011]. Meanwhile, large scale multimedia data sets collected from the Internet often inevitably contain noise and outliers. Therefore a robust version of matrix factorization is demanded to learn salient concepts of images. For this reason, $\ell_{2,1}$-norm loss function [Ding *et al.*, 2006] is introduced to design a robust matrix factorization. $\ell_{2,1}$-norm is defined for a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in \mathbb{R}^{p\times n}$ as

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{n} \|\mathbf{a}_i\|_2. \tag{5}$$

Compared with ℓ_2-norm, $\ell_{2,1}$-norm does not square the reconstruction error of each sample $\mathbf{a}_i$. Therefore, the objective function in Eq. (5) is expected to be robust to noise and outliers.

Let $\mathbf{X}$ be a set of m-dimensional image features, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$, the proposed robust matrix factorization can be formulated as

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_{2,1} = \sum_{j=1}^{n} \|\mathbf{x}_j - \mathbf{U}\mathbf{v}_j\|_2\ , \tag{6}$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k] \in \mathbb{R}^{m\times k}$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \in \mathbb{R}^{k\times n}$, and k is the length of the hash codes. By robust matrix factorization, each image feature $\mathbf{x}_j$ is approximated by a linear combination of the columns of $\mathbf{U}$, weighted by the components of $\mathbf{v}_j$. Then, $\mathbf{U}$ can be regarded as containing some semantic concepts and each image can be regarded as the linear combination of those concepts.

Note that in general, the cost function of $\ell_{2,1}$-norm form in Eq. (6) is harder to solve than the ℓ_2-norm form. Then we rewrite Eq. (6) as

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_{2,1} = Tr\left\{(\mathbf{X} - \mathbf{U}\mathbf{V})\,\mathbf{D}_X(\mathbf{X} - \mathbf{U}\mathbf{V})^T\right\}, \tag{7}$$

where $\mathbf{D}_X$ is a diagonal matrix with the j-th diagonal element given by

$$(\mathbf{D}_X)_{jj} = 1/\|\mathbf{x}_j - \mathbf{U}\mathbf{v}_j\|_2. \tag{8}$$

Therefore, the overall objective function for image concept modeling is

$$\min_{\mathbf{U},\mathbf{V}} \mathcal{L}_I = Tr\left\{(\mathbf{X} - \mathbf{U}\mathbf{V})\,\mathbf{D}_X(\mathbf{X} - \mathbf{U}\mathbf{V})^T\right\}. \tag{9}$$

By minimizing Eq. (9), the latent semantic concepts and the corresponding coefficients for images will be learned simultaneously.

2.3 Semantic Correlation Matching

If multimodal data points have the same semantics, they are expected to have a certain common latent space. For example, there are two semantic topics of text in Figure 1 (a) which are the natural landscape and cultural landscape as shown in Figure 1 (c). The relevant image in Figure 1 (b) also has some semantic concepts such as architectures, mountains, sky and sea as shown in Figure 1 (d). The topics of text actually have correlativity with the concepts of the image. For instance, natural landscape topic of text is related to the image concepts mountains, sky and sea. Therefore a topic in a text can be described by several concepts in an image. Accordingly, the cross-correlation for image and text can be formulated as

$$\min_{\mathbf{P}} \mathcal{L}_C = \|\mathbf{H} - \mathbf{P}\mathbf{V}\|_F^2 \ , \tag{10}$$

where $\mathbf{P} \in \mathbb{R}^{k\times k}$ is the correlation matrix between images and texts. By Eq. (10), the coefficients of image can be transformed into hash codes. Then, a text and image pair has the identical hash code which facilities the cross-media retrieval.

2.4 Overall Objective Function

The overall objective function, combining the semantic topic modeling for text given in Eq. (4), the robust matrix factorization for image given in Eq. (9), and the cross-correlation between the latent semantic spaces of image and text given in Eq. (10), is written as below.

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{H},\mathbf{U},\mathbf{V},\mathbf{P}} \mathcal{L} = \lambda\mathcal{L}_T + (1-\lambda)\,\mathcal{L}_I + \mu\mathcal{L}_C + \gamma R(\mathbf{F},\mathbf{U},\mathbf{V},\mathbf{P}) \\ &s.t.\ h_{ij} \in \{0,1\}\,, \sum_{i=1}^{k} h_{ij} = c\ , \end{aligned} \tag{11}$$

where λ, μ and γ are tradeoff parameters, and $R(\cdot) = \|\cdot\|_F^2$ is the regularization term to avoid overfitting.

2.5 Optimization Algorithm

The optimization problem in Eq. (11) can be solved by updating the following steps iteratively until convergency or the preset maximum number of iterations is reached.

1. Fix $\mathbf{H}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$, let the derivative of $\mathcal{L}$ with respect to $\mathbf{F}$ equals to zero, then we obtain

$$\mathbf{F} = \mathbf{Y}\mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T + \gamma/\lambda\mathbf{I}\right)^{-1}, \tag{12}$$

where $\mathbf{I}$ is the identity matrix.

2. Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{V}$, and $\mathbf{P}$, let the derivative of $\mathcal{L}$ with respect to $\mathbf{U}$ equals to zero, then we obtain

$$\mathbf{U} = \mathbf{X}\mathbf{D}_X\mathbf{V}^T\left(\mathbf{V}\mathbf{D}_X\mathbf{V}^T + \gamma/\left(1-\lambda\right)\mathbf{I}\right)^{-1}. \tag{13}$$

3. Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{U}$, and $\mathbf{V}$, let the derivative of $\mathcal{L}$ with respect to $\mathbf{P}$ equals to zero, then we obtain

$$\mathbf{P} = \mathbf{H}\mathbf{V}^T\left(\mathbf{V}\mathbf{V}^T + \gamma/\mu\mathbf{I}\right)^{-1}. \tag{14}$$

4. Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{U}$, and $\mathbf{P}$, let the derivative of $\mathcal{L}$ with respect to $\mathbf{V}$ equals to zero, then we obtain

$$\mathbf{v}_j = \left[\left(1-\lambda\right)\left(\mathbf{D}_X\right)_{jj}\mathbf{U}^T\mathbf{U} + \mu\mathbf{P}^T\mathbf{P} + \gamma\mathbf{I}\right]^{-1}\mathbf{q}_j\ , \tag{15}$$

where $\mathbf{q}_j$ is the j-th column of $\left(1-\lambda\right)\mathbf{U}^T\mathbf{X}\mathbf{D}_X + \mu\mathbf{P}^T\mathbf{H}$.

5. Fix $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$, we have the following equation to calculate $\mathbf{H}$

$$\begin{aligned} &\min_{\mathbf{H}} \sum_{j=1}^{n}\left(\sum_{i=1}^{k} h_{ij}\left\|\mathbf{y}_j - \mathbf{f}_i\right\|_2^2 + \tfrac{\mu}{\lambda}\left\|\mathbf{h}_j - \mathbf{P}\mathbf{v}_j\right\|_2^2\right) \\ &s.t.\ h_{ij} \in \{0,1\}\,, \sum_{i=1}^{k} h_{ij} = c\ . \end{aligned} \tag{16}$$

Eq. (16) can be solved by calculating $\mathbf{h}_j$ one by one independently. That is, we need to solve the following problem for $\mathbf{h}_j$

$$\begin{aligned} &\min_{\mathbf{h}_j} \sum_{i=1}^{k} h_{ij}\|\mathbf{y}_j - \mathbf{f}_i\|_2^2 + \tfrac{\mu}{\lambda}\left\|\mathbf{h}_j - \mathbf{P}\mathbf{v}_j\right\|_2^2 \\ &= \sum_{i=1}^{k} h_{ij}\|\mathbf{y}_j - \mathbf{f}_i\|_2^2 + \tfrac{\mu}{\lambda}\left(\mathbf{h}_j^T\mathbf{h}_j - 2\mathbf{h}_j^T\mathbf{P}\mathbf{v}_j + \mathbf{v}_j^T\mathbf{P}^T\mathbf{P}\mathbf{v}_j\right) \\ &= \sum_{i=1}^{k} h_{ij}\left[\|\mathbf{y}_j - \mathbf{f}_i\|_2^2 - \tfrac{2\mu}{\lambda}\left(\mathbf{P}\mathbf{v}_j\right)_i\right] + C \\ &s.t.\ h_{ij} \in \{0,1\}\,, \sum_{i=1}^{k} h_{ij} = c\ , \end{aligned} \tag{17}$$

where $\left(\mathbf{P}\mathbf{v}_j\right)_i$ is the i-th element of vector $\mathbf{P}\mathbf{v}_j$ and C is a constant. To minimize Eq. (17), we first rank $\|\mathbf{y}_j - \mathbf{f}_i\|_2^2 - 2\mu/\lambda(\mathbf{P}\mathbf{v}_j)_i$, $i = 1, 2, \cdots, k$, with the order small to large. Then let $h_{ij} = 1$ if $\|\mathbf{y}_j - \mathbf{f}_i\|_2^2 - 2\mu/\lambda(\mathbf{P}\mathbf{v}_j)_i$ belongs to the top c minimum values in the ranking list and otherwise $h_{ij} = 0$.

2.6 Out-of-Sample Extension

Text

Let $\mathbf{y}^t \in \mathbb{R}^{d\times 1}$ be the query text feature, then its hash code $\mathbf{h}^t$ can be obtained by

$$\begin{aligned} &\min_{\mathbf{h}^t} \sum_{i=1}^{k} h_i^t\left\|\mathbf{y}^t - \mathbf{f}_i\right\|_2^2 \\ &s.t.\ h_i^t \in \{0,1\}\,, \sum_{i=1}^{k} h_i^t = c\ , \end{aligned} \tag{18}$$

where h_i^t is the i-th element of vector $\mathbf{h}^t$. It can be solved by ranking $\|\mathbf{y}^t - \mathbf{f}_i\|_2^2$, $i = 1, 2, \cdots, k$, with the order small to large. And then let $h_i^t = 1$ if $\|\mathbf{y}^t - \mathbf{f}_i\|_2^2$ belongs to the top c minimum values in the ranking list and otherwise $h_i^t = 0$.

Image

Let $\mathbf{x}^t \in {}^{m\times 1}$ be the image query feature, then its hash code $\mathbf{h}^t$ can be obtained by first embedding it to the semantic concept space $\mathbf{U}$ by

$$\mathbf{v}^t = \left[\mathbf{U}^T\mathbf{U} + \gamma/\left(1-\lambda\right)\mathbf{I}\right]^{-1}\mathbf{U}^T\mathbf{x}^t. \tag{19}$$

Then transform $\mathbf{v}^t$ to the hash space by

$$\mathbf{h}^t = sign\left[\mathbf{P}\mathbf{v}^t - median(\mathbf{P}\mathbf{v}^t)\right], \tag{20}$$

where $sign(\cdot)$ denotes the sign function and $median(\cdot)$ denotes the median function.

We summarize the procedures for STMH in Algorithm 1.

Algorithm 1 Semantic Topic Multimodal Hashing

Training:

Input: Images $\mathbf{X}$, texts $\mathbf{Y}$, parameters λ, μ, and γ, bit length k.

Output: Hash codes $\mathbf{H}$, matrices $\mathbf{F}$, $\mathbf{U}$, and $\mathbf{P}$.

Procedure:

1. Initialize $\mathbf{F}$, $\mathbf{H}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$;
2. Repeat
   2.1 Compute $\mathbf{D}_X$ with Eq. (8);
   2.2 Fix $\mathbf{H}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$, update $\mathbf{F}$ with Eq. (12);
   2.3 Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{V}$, and $\mathbf{P}$, update $\mathbf{U}$ with Eq. (13);
   2.4 Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{U}$, and $\mathbf{V}$, update $\mathbf{P}$ with Eq. (14);
   2.5 Fix $\mathbf{F}$, $\mathbf{H}$, $\mathbf{U}$, and $\mathbf{P}$, update $\mathbf{V}$ with Eq. (15);
   2.6 Fix $\mathbf{F}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$, update $\mathbf{H}$ by solving Eq. (17).
   until convergency.
3. Return $\mathbf{H}$, $\mathbf{F}$, $\mathbf{U}$, and $\mathbf{P}$.

Testing:

Input: Image $\mathbf{x}^t$ or text $\mathbf{y}^t$, matrices $\mathbf{F}$, $\mathbf{U}$, and $\mathbf{P}$.

Output: Hash code $\mathbf{h}^t$.

Procedure:

Text: Get the hash code $\mathbf{h}^t$ by solving Eq. (18).

Image: Get the hash code $\mathbf{h}^t$ with Eqs. (19) and (20).

2.7 Computational Complexity Analysis

The computational complexity for training STMH is $O\left(\left(k^3+dk+mk\right)nt\right)$, where t is the number of iterations. As k, d, m and $t \ll n$, the training complexity is linear to the training data size. In the search phase, the complexity is constant with $O\left(dk\right)$ for a text query and $O\left(mk\right)$ for an image query. In a word, the time complexity for training STMH is linear to n and is constant for testing, which is really scalable for large-scale data sets.

2.8 Multiple Modalities Extension

The extension for STMH in Eq. (11) from bimodal to multiple modalities is quite easy and direct by using semantic topic modeling for one chosen modality and robust matrix factorization for others.

$$\begin{aligned}\min_{\mathbf{F},\mathbf{H},\mathbf{U}_t,\mathbf{V}_t,\mathbf{P}_t} \mathcal{L} = &\left(1-\sum_t \lambda_t\right)\left[\sum_{j=1}^{n}\sum_{i=1}^{k} h_{ij}\|\mathbf{y}_j-\mathbf{f}_i\|_2^2 + Tr\left(\mathbf{F}\mathbf{W}\mathbf{F}^T\right)\right] \\ &+ \sum_t \lambda_t \|\mathbf{X}_t - \mathbf{U}_t\mathbf{V}_t\|_{2,1} + \mu \sum_t \|\mathbf{H}-\mathbf{P}_t\mathbf{V}_t\|_F^2 \\ &+ \gamma\left[\sum_t R(\mathbf{P}_t,\mathbf{U}_t,\mathbf{V}_t) + R(\mathbf{F})\right] \\ s.t.\ h_{ij} \in \{0,1\}, &\sum_{i=1}^{k} h_{ij} = c\ .\end{aligned} \tag{21}$$

It is straightforward to adapt Algorithm 1 presented above to solve Eq. (21).

3 Experiments

To evaluate the performance of the proposed STMH, we conduct comparison experiments with several state-of-art methods including CMSSH [Bronstein *et al.*, 2010], CVH [Kumar and Udupa, 2011], LCMH [Zhu *et al.*, 2013], IMH [Song *et al.*, 2013], CMFH [Ding *et al.*, 2014], and LSSH [Zhou *et al.*, 2014], on two real-world datasets, i.e., Wiki[1] and NUS-WIDE[2] for cross-media similarity search. The retrieval performance is evaluated by mean of average precision (mAP), *recall-precision*, and *topN-precision*. Details about data sets processing and evaluation metrics can be referred to [Zhou *et al.*, 2014].

3.1 Experimental Settings

CMSSH, IMH and LSSH require too much computational costs that are quite difficult to learn hash functions on NUS-WIDE with all data. Thus, we randomly select 5000 instances from data set for these methods to train hash functions and then apply the trained hash functions to the other instances in data set to generate hash codes for them as [Song *et al.*, 2013] did.

For all the comparison algorithms except LCMH, the codes are kindly provided by the authors. We implemented LCMH as the code is not publicly available. The parameters for all the comparison methods are tuned according to the corresponding literatures. When comparing with the baseline methods, we use the parameter settings, $\lambda = 0.5$, $\mu = 0.001$, and $\gamma = 10^{-4}$ for STMH.

3.2 Results and Discussions

The mAP values for STMH and six baseline methods are reported in Figure 2. The *recall-precision* and *topN-precision* curves are plotted in Figure 3 and Figure 4 respectively.

Results on Wiki

The Wiki data set is separated into two parts, with 2173 pairs for training and 693 pairs for testing. It can be observed that STMH achieves the best performance than baseline methods on text-query-image similarity search task. And the performance for STMH on image-query-text task although does not always achieve the best performance, it achieves comparable performance to the best. As shown in Figure 4, we can find that STMH achieves better performance on the top 400 retrieved instances. In retrieval system, the top retrieved instances often need high accuracy as users care more about the front instance in the retrieved list. From this point of view, STMH can achieve relatively good performance on image-query-text task.

Results on NUS-WIDE

We randomly choose 1K images with their tags to serve as the test set and the rest images and tags are serving as the training set in NUS-WIDE data set. As the results show, STMH outperforms baseline methods significantly which verifies that it can better model the complex structure of large-scale data set by their latent semantic topics and reduce the semantic gap between heterogeneous data compared with state-of-the-art methods. Furthermore, STMH achieves higher mAP values with longer codes. This is reasonable as longer hash codes can encode more semantic information and therefore can improve the retrieval accuracy. As NUS-WIDE data set is quite similar to real-world scenario, experimental results show that STMH can handle large-scale multimodal similarity search problem.

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

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

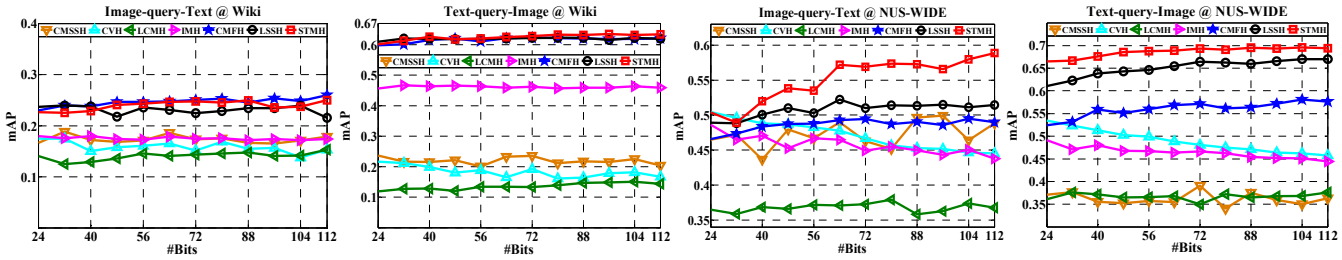

Figure 2: mAP on Wiki and NUS-WIDE with different code lengths.

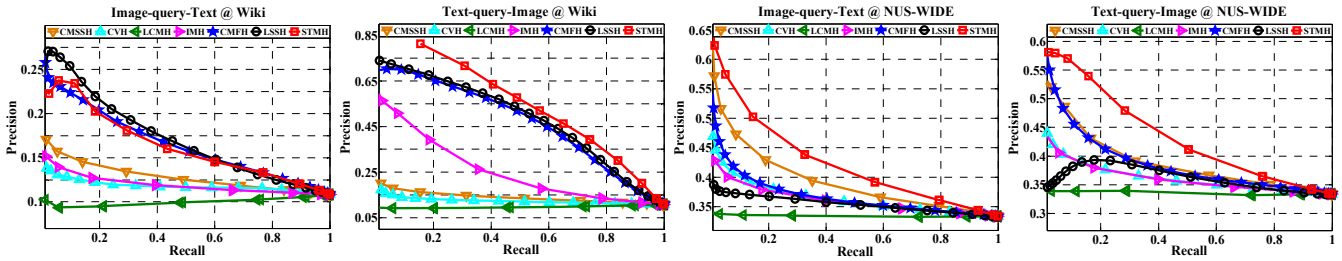

Figure 3: *Recall-Precision* curves on Wiki and NUS-WIDE with 32 bits code length.

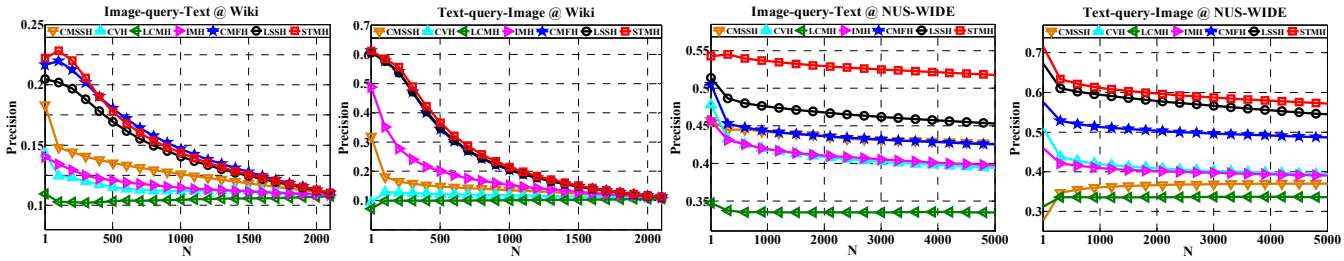

Figure 4: *topN-Precision* curves on Wiki and NUS-WIDE with 64 bits code length.

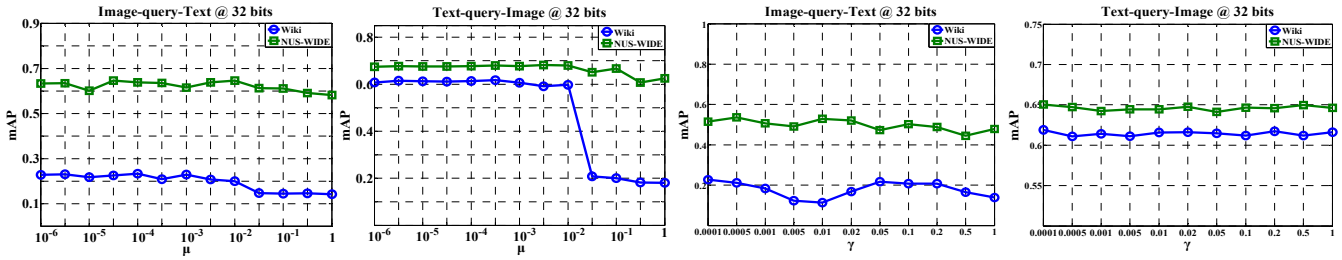

Figure 5: Parameter sensitivity analysis

3.3 Parameter Sensitivity Analysis

The empirical analysis on parameter sensitivity is given in this section. The analysis is conducted for one parameter by varying its value while fixing the other parameters.

The parameter μ controls the connection of latent semantic spaces between image and text. If it is too large, the strong connection will affect the learning of latent semantic topics of texts and images. However, if it is too small, the connection between different modalities is weak which will result in poor performance for cross-modality similarity search. It is reasonable to choose proper value of μ from the range of $[0.0001, 0.01]$.

The parameter γ controls the complexity of the model. The model is over-fitted with too small value while under-fitted with too large value. It can be observed from Figure 5 that STMH can achieve stable performance under a wide range of γ. Usually, it can be chosen from the range of $[0.0001, 0.1]$.

From the above analysis, we can reach the conclusion that STMH can achieve stable performance under a wide range of parameter values.

4 Conclusions

In this paper, we presented a novel multimodal hashing method, referred to as semantic topic multimodal hashing (STMH), for large-scale cross-media similarity search. Specifically, STMH models text as multiple semantic topics and image as latent semantic concepts and learns the relationship of text and image in their latent semantic spaces. Then, each bit of unified hash code can be generated directly by figuring out whether a topic or concept is contained in a text

or an image. By maintaining the discrete nature of hashing, STMH is more suitable for hashing learning scheme and can obtain better retrieval performance. Comparative studies on two bench-mark datasets show that STMH outperforms the state-of-the-art multimodal hashing methods. In future work, we will explore more efficient semantic information to further improve the performance of STMH.

Acknowledgments

This paper was supported partially by the National Natural Science Foundation of China (Grant Nos. 61125204, 61432014, 61472304, and 61172146), the Fundamental Research Funds for the Central Universities (Grant Nos. BDZ021403 and JB149901), the Program for Changjiang Scholars and Innovative Research Team in University of China (No. IRT13088), the Shaanxi Innovative Research Team for Key Science and Technology (No. 2012KCT-02) and the Project Funded by China Postdoctoral Science Foundation (No. 2014M562378).

References

[Bedi *et al.*, 2007] Punam Bedi, Harmeet Kaur, and Sudeep Marwaha. Trust based recommender system for semantic web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2677–2682, Hyderabad, India, January 2007. AAAI Press.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Blei, 2012] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[Bronstein *et al.*, 2010] Michael M. Bronstein, Alexander M. Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using simi-larity-sensitive hashing. In *Proceedings of the 23rd IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, San Francisco, CA, USA, June 2010. IEEE Computer Society.

[Cai *et al.*, 2011] Deng Cai, Xiaofei He, Jiawei Han, and T. S. Huang. Graph regularized nonnegative matrix factoriza-tion for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[Ding *et al.*, 2006] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R_1-PCA: rotational invariant L_1-norm principal component analysis for robust subspace fac-torization. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 281–288, Pittsburgh, Pennsylvania, USA, June 2006. ACM.

[Ding *et al.*, 2014] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of the 27th IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2083–2090, Columbus, OH, USA, June 2014. IEEE Computer Society.

[Kong *et al.*, 2011] Deguang Kong, Chris Ding, and Heng Huanga. Robust nonnegative matrix factorization using $L_{2,1}$-norm. In *Proceedings of the 20th ACM International Conference on Information and knowledge management*, pages 673–682, Glasgow, United Kingdom, October 2011. ACM.

[Kumar and Udupa, 2011] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1360–1367, Barcelona, Catalonia, Spain, July 2011. AAAI Press.

[Lee and Seung, 1999] Daniel D. Lee and Sebastian H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Song *et al.*, 2013] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Hengtao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 19th International Conference on Management of Data*, pages 785–796, Ahmedabad, India, December 2013. Computer Society of India.

[Weiss *et al.*, 2009] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, pages 1753–1760, Vancouver, British Columbia, Canada, December 2009. Curran Associates.

[Weston *et al.*, 2011] Jason Weston, Samy Bengio, and Ni colas Usunier. Wsabie: scaling up to large vocabulary image annotation. In *Proceedings of the 22nd Interna-tional Joint Conference on Artificial Intelligence*, pages 2764–2770, Barcelona, Catalonia, Spain, July 2011. AAAI Press.

[Zhai *et al.*, 2013] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, and Wen Gao. Parametric local multimodal hashing for cross-view similarity search. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2754–2760, Beijing, China, August 2013. AAAI Press.

[Zhou and Tao, 2011] Tianyi Zhou and Dacheng Tao. Godec: randomized low-rank & sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning*, pages 33–40, Bellevue, Washington, USA, June 2011. ACM.

[Zhou *et al.*, 2014] Jile Zhou, Guiguang Ding, and Yuhen Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–424, Gold Coast, QLD, Australia, July 2014. ACM.

[Zhu *et al.*, 2013] X. Zhu, Z. Huang, H. Shen, Z. Huang, and X. Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM Multimedia Conference*, pages 143–152, Barcelona, Spain, October 2013. ACM.