11-2017

Unsupervised topic hypergraph hashing for efficient mobile image retrieval

Lei ZHU
Singapore Management University

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Liang XIE
Wuhan University of Technology

Zhiyong CHENG
Singapore Management University, zy.cheng.2011@phdis.smu.edu.sg

DOI: https://doi.org/10.1109/TCYB.2016.2591068

Citation

ZHU, Lei; SHEN, Jialie; XIE, Liang; and CHENG, Zhiyong. Unsupervised topic hypergraph hashing for efficient mobile image retrieval. (2017). *IEEE Transactions on Cybernetics*. 47, (11), 3941-3954. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3537

Published in IEEE Transactions on Cybernetics, 2016 Oct, Advanced Online.
http://doi.org/10.1109/TCYB.2016.2591068

Unsupervised Topic Hypergraph Hashing for Efficient Mobile Image Retrieval

Lei Zhu, Jialie Shen, Liang Xie, and Zhiyong Cheng

***Abstract*—Hashing compresses high-dimensional features into compact binary codes. It is one of the promising techniques to support efficient mobile image retrieval, due to its low data transmission cost and fast retrieval response. However, most of existing hashing strategies simply rely on low-level features. Thus, they may generate hashing codes with limited discriminative capability. Moreover, many of them fail to exploit complex and high-order semantic correlations that inherently exist among images. Motivated by these observations, we propose a novel unsupervised hashing scheme, called topic hypergraph hashing (THH), to address the limitations. THH effectively mitigates the semantic shortage of hashing codes by exploiting auxiliary texts around images. In our method, relations between images and semantic topics are first discovered via robust collective non-negative matrix factorization. Afterwards, a unified topic hypergraph, where images and topics are represented with independent vertices and hyperedges, respectively, is constructed to model inherent high-order semantic correlations of images. Finally, hashing codes and functions are learned by simultaneously enforcing semantic consistence and preserving the discovered semantic relations. Experiments on publicly available datasets demonstrate that THH can achieve superior performance compared with several state-of-the-art methods, and it is more suitable for mobile image retrieval.**

***Index Terms*—High-order semantic correlations, mobile image retrieval, topic hypergraph hashing (THH).**

I. Introduction

THE RECENT decades have witnessed rapid growth of social image websites (e.g., Flickr[1] and Instagram).[2] As a result, huge amount of images have been recorded and shared on the Web. On the other hand, with fast popularity of smart mobile devices, mobile image retrieval [1] is gaining in importance due to many potential applications, such as landmark retrieval [2], document image retrieval [3], product retrieval [4], etc. The most popular and naive approach to support mobile image retrieval is text-based retrieval, where users are required to type text keywords as query. However, it is very time-consuming and inconvenient on mobile devices. Thus, content-based mobile image retrieval (CBMIR) [5], where only visual images are uploaded as queries, becomes a popular and convenient retrieval paradigm.

Different from other computing platforms, mobile devices generally have limited computational power, memory, and battery capacity. Hence, most practical CBMIR systems apply a client-server architecture: visual queries are uploaded from mobile end and sent to powerful server. Time-consuming retrieval process can be efficiently completed with rich computing resources. However, mobile devices are usually located in a context with limited wireless network bandwidth. Therefore, to support efficient and effective CBMIR, the transmitted query data should be both compact and semantically discriminative.

Hashing [6]–[18] can be applied as an effective technique for CBMIR. The core idea is to transform high-dimensional data into compact binary codes, based on which similarities of images are measured with Hamming distance. With hashing as underlying indexing, the storage occupation of query in mobile memory can be greatly reduced and the query data can be efficiently transmitted. More importantly, hashing can binarize the visual data for both query and database images. Retrieval process can be completed with simple but efficient bit operations. Unfortunately, existing hashing techniques generally suffer from two major limitations when being directly applied to CBMIR [19].

1) CBMIR is only based on visual queries. Therefore, most hashing schemes employ only low-level visual features. Due to the well-known semantic gap, they cannot represent high-level semantics of images effectively. On the other hand, it is common that real world database images (e.g., those from Flickr or Wiki)[3] for CBMIR are accompanied with informative tags or textual descriptions. Cross-modal/media hashing (CMH) [20]–[22] can potentially exploit these resources by projecting visual contents and texts into common subspace. However, the main aim of CMH is to support efficient retrieval across image and text. Therefore, CMH generally treats the involved visual and textual information equally. Consequently, the semantics of images shared in the common Hamming space are very limited.

Manuscript received December 17, 2015; revised March 31, 2016; accepted June 27, 2016. This paper was recommended by Associate Editor H. Wang. *(Corresponding author: Jialie Shen).*

L. Zhu and Z. Cheng are with the School of Information Systems, Singapore Management University, Singapore.

J. Shen is with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, NE1 8ST, U.K. (e-mail: jialie@gmail.com).

L. Xie is with the School of Science, Wuhan University of Technology, Wuhan 430070, China.

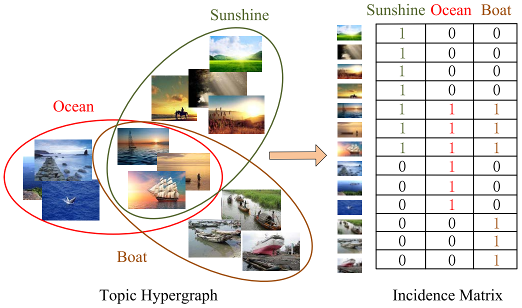
Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2591068

[1] https://www.flickr.com/

[2] https://instagram.com/

[3] https://www.wikipedia.org/



Sunshine	Ocean	Boat
1	0	0
1	0	0
1	0	0
1	0	0
1	1	1
1	1	1
1	1	1
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

Fig. 1. Semantic correlations of images are inherently high-order. This paper describes them with a unified THG and preserves them in hashing codes. The above figure presents a typical THG and its corresponding incidence matrix.

2) Most existing hashing strategies perform learning on pairwise visual or textual similarities of images. Thus, they cannot capture high-order semantic correlations that inherently exist among real Web images (as shown in Fig. 1). It is common that a single image represents more than one semantic topic, and images may jointly share several common semantic topics. Besides, even in Hamming space, we can find that correlations of images are high-order, by assuming that each hashing bit represents a latent relation (hyperedge). Hence, it is essential to discover high-order semantic correlations of images, and more importantly, preserve them in hashing codes.

While jointly modeling auxiliary media types can enhance semantic understanding, promising results are achieved in image retrieval [23]–[25]. For example, [23] treats all of the affiliated media objects of an image as a whole. Hierarchical manifolds for multimedia contents are leveraged for content-based cross-modal image retrieval. Similarly, Rasiwasia *et al.* [24] investigated the effects of explicitly modeling relations of images and the associated texts. Their experiments validate that modality correlation can benefit image retrieval. Other examples include [26] and [27]. Yang *et al.* [26] harvested informative relevance feedbacks regrading images to mitigate the semantic gap of low-level features using a semi-supervised rank framework. Yan *et al.* [27] explored surrounding texts to improve Web image clustering by mining semantic correlations between text words and images. Based on the clustering results, diverse images are retrieved with conditional Markov random walk. The success of these works motivates us to develop image hashing technique by exploiting the assistance of auxiliary texts.

In this paper, we propose a novel unsupervised hashing scheme, called topic hypergraph hashing (THH), to support effective and efficient CBMIR. The key idea is to extract valuable semantics from auxiliary texts to assist image hashing. THH first discovers image-topic relations via robust collective non-negative matrix factorization. Then, it constructs a unified topic hypergraph (THG) to model high-order semantic correlations of images. Unlike existing hypergraphs built on low-level features [28], THG is constructed with informative semantics by modeling images as vertices and latent topics shared among images and texts as hyperedges. Finally, hashing learning is carried out to preserve semantics into binary hashing codes and enhance the discriminative representation capability. THH has several desirable advantages which can effectively facilitate mobile image retrieval.

1) THH exploits visual and textual contents in offline learning but requires only visual image as online query. This design undoubtedly provides great convenience for mobile users, since typing text keywords on mobile ends is harder compared with a simple photo snapping.
2) THH captures the characteristic of mobile image retrieval that database images are usually associated with noisy but informative texts. It effectively extracts valuable semantics and compresses them into hashing codes. With this design, the discriminative capability of hashing codes can be enhanced. As revealed in experiments (in Section VI), THH can achieve better performance with even less hashing codes. This promising property of THH can facilitate effective mobile image retrieval even when wireless bandwidth is limited. Further, THH is a linear method. The online hashing process can be efficiently implemented with linear operations. This desirable property can facilitate practical application of mobile image retrieval where computational resources are limited.

The key contributions of this paper are summarized as follows.

1) Different from existing hypergraphs based on pure low-level features, a novel THG is constructed with the semantic assistance of auxiliary texts around images. THG effectively models the high-order semantic correlations that inherently exist among real Web images.
2) Hashing learning is performed by effectively preserving the high-order semantic relations of images into binary codes. The whole process is integrated into an unsupervised learning framework which enriches semantics of hashing codes without any manual labels.
3) Experiments on publicly available datasets highlight various advantages of THH and demonstrate that it significantly outperforms several state-of-the-art hashing methods from various perspectives.

The rest of this paper is structured as follows. Section II introduces related work. System overview of THH-based CBMIR system is illustrated in Section III. Details about THH are introduced in Section IV. Experimental configuration is presented in Section V. In Section VI, we give experimental results and analysis. Section VII concludes this paper with a detailed summary and future work.

II. Related Work

Due to the limited space here, only the work highly related to this paper is introduced. In particular, we present a short literature review on mobile image retrieval, unsupervised hashing, and hypergraph learning for image retrieval.

A. Mobile Image Retrieval

Since transmitting an entire query photo via wireless network is time-consuming, many existing techniques on mobile

TABLE I
COMPARISONS OF REPRESENTATIVE HASHING SCHEMES AND THE PROPOSED THH

Methods	Query	Learning Feature	Learning Space	Semantic Enhancement	Generality	CBMIR
Visual-words based Hashing	Visual	Visual	Visual	No	No	Yes
Uni-modal Hashing	Visual	Visual	Visual	No	Yes	Yes
Multi-modal Hashing	Visual+Text	Visual+Text	Multi-modal fused	Yes	Yes	No
Cross-modal Hashing	Visual or Text	Visual+Text	Common	Limited	Yes	Partly
The proposed THH	Visual	Visual+Text	Text-enhanced visual	Yes	Yes	Yes

image retrieval focus on creating compact feature descriptors for raw query. Various feature compression strategies are developed. For example, transform coding and location histogram coding are proposed in [29] and [30], respectively, to compress local features. Besides, there are also methods which are designed to compress features of particular image types. Ji *et al.* [2] presented a location discriminative vocabulary coding by exploiting the pervasive location context to obtain compact visual landmark descriptor. He *et al.* [3] designed bag-of-hashing-bits (BoHB) for mobile product retrieval. BoHB encodes the local feature into binary bits by leveraging techniques such as multitable indexing, multibucket probing, and bit reuse. Duan *et al.* [4] developed memory-light document indexing with codebook-free scalable cascaded hashing.

Although existing techniques for mobile image retrieval can achieve promising results, they are specifically designed for compressing visual-words-based features or particular image types. Hence, they cannot be directly applied to general feature or image types. Moreover, they are designed based on low-level features with limited semantic discriminative capability. This disadvantage further limits the hashing performance.

B. *Unsupervised Hashing*

Distinguished from supervised [31] and semi-supervised hashing methods [32], unsupervised hashing transforms the original feature into binary codes without any semantic labels. This desirable advantage can effectively cope with the practical CBMIR, where semantic labels are quite scarce and expensive to obtain. Generally, existing unsupervised hashing techniques can be categorized into three major families: 1) uni-modal hashing (UMH); 2) cross-modal hashing (CMH); and 3) multimodal hashing (MMH).

State-of-the-art UMH methods include: locality-sensitive hashing (LSH) [15], spectral hashing (SPH) [8], PCA hashing (PCAH) [33], binarised-LSI (LSI) [34], spline regression hashing (SRH) [35], self-taught hashing (STH) [36], anchor graph hashing (AGH) [37], iterative quantization (ITQ) [16], supervised hashing with pseudo labels [38], etc. Although UMH demonstrates promising performance, it suffers from several limitations. The most significant one is that UMH only takes the features from visual modality into account. Due to the well-known semantic gap, image similarity characterized by only low-level visual feature may not be comprehensively enough to describe the semantics of images. Consequently, UMH learns the hashing codes with limited discriminative capability.

CMH discovers latent modality correlations and transforms heterogeneous modalities into the common Hamming space, where similarities are quickly computed to return retrieval results. Typical CMH methods include: cross-view hashing (CVH) [20], intermedia hashing (IMH) [22], and collective matrix factorization hashing (CMFH) [39]. It has been reported in [22] that, due to textual modality embedding, the shared space can possess more semantics than original low-level visual space. Therefore, CMH may improve the CBMIR performance. However, the main aim of various CMH approaches is to enable fast multimedia retrieval across heterogeneous modalities. Principally, it requires that each involved modality contributes equally to image retrieval. This mandatory correlation limits the semantics involved in the shared common Hamming space.

MMH compresses multimodal features into a unified binary codes. Composite hashing with multiple information sources (CHMIS) [21] is one of the pioneering works. It learns hashing codes by incorporating the features from different information sources. However, CHMIS just postintegrates linear output of features and fails to fully exploit the correlations of them. Kim *et al.* [40] presented multiview SPH using sequential projection learning [33] to extend SPH [8] to handle multiple image representations. Song *et al.* [41] developed multiple feature hashing (MFH). By using the learned hashing hyper plane, MFH concatenates all features into a single vector and then maps it into binary hashing codes. Zou *et al.* [42] designed kernelized MFH to learn compact fingerprint by integrating advantages of nonlinear kernel mapping and the complements of multiple features. Liu *et al.* [43] proposed a multiview alignment hashing by aligning multimodal features into a joint hashing space. Due to multifeature fusion, MMH can achieve better performance than UMH and CMH. However, it requires all modalities at both stages of offline learning and online hashing. Due to this constraint, MMH cannot meet the requirement of CBMIR in practical applications, where only visual image is uploaded as query.

Table I summarizes key characteristics of several representative hashing schemes. From this table, we can easily find that, THH not only can support effective CBMIR, but also can leverage the associated texts to enrich the semantics of projected hashing codes. Moreover, it is independent on basis features and thus has desirable generalization capability.

C. *Hypergraph Learning for Image Retrieval*

Hypergraph is an extension of graph [44]–[47] which models pairwise relations of samples. For its advantage on

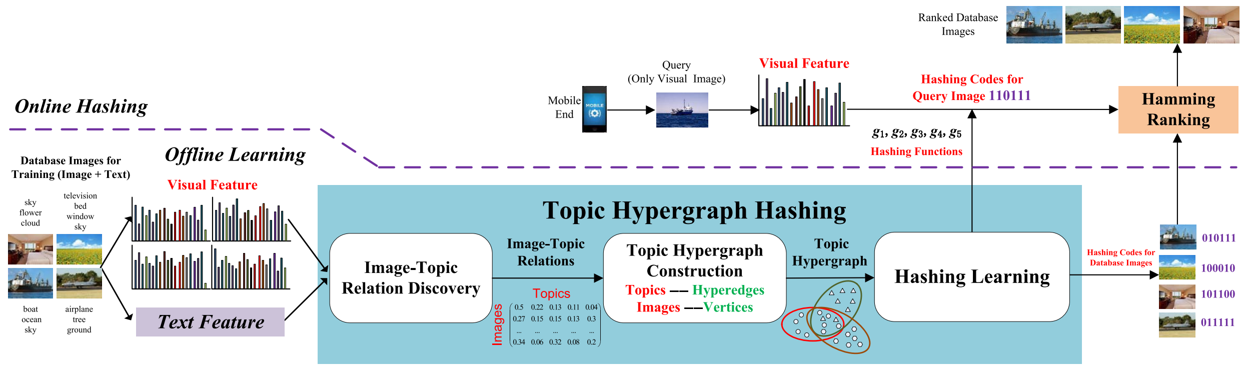


Fig. 2. Framework of the THH-based CBMIR system. This figure is best viewed with pdf magnification.

modeling complex data, hypergraph has attracted great attentions in image retrieval. One of typical examples is the work in [48]. In this paper, image retrieval is formulated in a probabilistic hypergraph (PHG) framework. Images are taken as vertices and a hyperedge is comprised of a centroid vertex and its several nearest neighbors. A vertex is assigned to hyperedge in a probabilistic way. Image retrieval process is transformed as hypergraph ranking. Gao *et al.* [49] applied hypergraph to social image retrieval by jointly learning visual-text relevance. In their constructed social image hypergraph, vertices represent images and hyperedges represent visual or textual terms. The weights of both visual elements and texts are learned adaptively in an iterative process. In [50], hypergraph is leveraged to solve the problem of view-based 3-D object retrieval. Each vertex is an object and a cluster of views constructs a hyperedge. K-means [51] is adopted to generate multiple overlapping hyperedges.

This paper proposes novel THG to solve large-scale CBMIR. Different from existing methods, in THG, images are determined as vertices and latent semantic topics are considered as hyperedges. With it, inherent high-order semantic relations of images are effectively modeled. And they are preserved in the hashing codes by THH to enhance semantic representation capability.

III. System Overview

As shown in Fig. 2, the system mainly consists of two key components: 1) offline learning and 2) online hashing.

1) *Offline Learning:* The aim of this process is to learn hashing functions which can map high-dimensional visual features of both query and database images into binary hashing codes. More specifically, offline learning can be further divided into four subsequent subprocesses: a) feature extraction; b) image-topic relation discovery; c) THG construction; and d) hashing learning. In the system, visual and textual features are first extracted from visual contents and the associated texts to represent images, respectively. Then, relations between images and topics are discovered by comprehensively considering visual and textual content distribution. Next, THG is constructed and high-order semantic relations of images are effectively modeled. Finally, hashing learning is performed by simultaneously preserving semantic relations into hashing codes and generating hashing functions.
2) *Online Hashing:* Query image is first submitted by user from mobile devices. Only visual feature of it is obtained by the same feature extraction pipeline conducted on database images. Then, it is projected into binary codes with the hashing functions learned from offline learning. Finally, the estimated similarity scores are computed with simple bit operations and ranked in descending order, and their corresponding database images are returned.

IV. Topic Hypergraph Hashing

This section provides the details of the proposed THH. First, we introduce notations and problem setting. Second, we give details of image-topic relation discovery. Third, we present THG construction. Fourth, we formulate the hashing learning and give an efficient solution. Finally, we summarize THH and give a computation complexity analysis.

A. Notations and Problem Setting

The transpose of matrix X is denoted as X^{T}. The inverse of a matrix X is denoted as X^{-1}. The trace operator on a matrix X is denoted as $\mathrm{Tr}(X)$. $||\cdot||_F$ denotes Frobenius norm. $\mathtt{sgn}(\cdot)$ is Sign function. I denotes identity matrix and $\mathbf{1}$ denotes a vector with all one elements. The corresponding dimensions of them can be inferred from the context.

Let $X_m = [x_1^m, \ldots, x_N^m] \in \mathbb{R}^{d_m \times N}$, $m = 1, 2$ denote features of database images extracted from visual contents and texts, respectively, $x_i^m = [x_{1i}^m, \ldots, x_{d_m i}^m]^{\mathrm{T}} \in \mathbb{R}^{d_m \times 1}$ denotes ith image feature from modality m, d_m denote the feature dimension, N is the number of database images. Denote hashing codes of database images to be learned as $Z = [z_1, \ldots, z_N] \in \mathbb{R}^{K \times N}$, and a group of hashing functions as $G = \{g_1, \ldots, g_K\}$, where $z_i = [z_{1i}, \ldots, z_{Ki}]^{\mathrm{T}} \in \mathbb{R}^{K \times 1}$ are the hashing codes of the ith image, each hashing function g_k is a mapping: $\mathbb{R}^{d_m} \mapsto \{0, 1\}$, $k = 1, \ldots, K$, K is the length of the hashing codes. The main notations used in the study are listed in Table II.

TABLE II
SUMMARY OF NOTATIONS

Symbols	Explanations
N	Number of database images
K	Number of topics and also hashing code length
S	Number of topics that images describe
X_m	$d_m \times N$ Feature representation in modality m
G	$\{g_k\}_{k=1}^{K}$ Group of hashing functions
Z	$K \times N$ Hashing codes of database images
B_m	$d_m \times K$ Topic matrix in modality m
T_m	$K \times N$ Image-topic relation matrix in modality m
T	$K \times N$ Shared image-topic relation matrix
α_m	Importance weight of modality m
THG	Topic hypergraph
V	$\{v_i\}_{i=1}^{N}$ Vertex set of THG
E	$\{e_k\}_{k=1}^{K}$ Hyperedge set of THG
W	$\{w(e_k)\}_{k=1}^{K}$ Weight set of THG
H	$K \times N$ Incidence matrix of THG
Δ	$N \times N$ Laplacian matrix of THG
D_v	$N \times N$ Diagonal matrix of the vertex degrees in THG
D_e	$K \times K$ Diagonal matrix of the hyperedge degrees in THG
D_w	$K \times K$ Diagonal matrix of the hyperedge weights in THG
P	$d_1 \times K$ Hashing projection matrix
μ, β, γ	Hyper-parameters

B. Image-Topic Relation Discovery

As shown in Fig. 1, real Web images are correlated with latently embedded semantic topics. To model high-order semantic relations of images, the first task is to detect semantic topics and discover the relations between images and topics. However, this task is challenging, because low-level visual features suffer from limited semantic discriminative capability. Fortunately, most database images for CBMIR are augmented with informative texts, such as tags, descriptions, image captions, etc. They generally carry out informative semantics and complementary discriminative information. Motivated by these important observations, this paper exploits auxiliary textual modality associated around the images to perform semantic assistance and address the image-topic relation discovery. The learning framework is based on non-negative matrix factorization [52]. Via comprehensively considering visual and textual distributions, the framework can simultaneously detect latent semantic topics and discover image-topic relation.

Generally, feature matrix of images can be represented as the product of two matrices. One is basis matrix and the other is coefficient matrix. By imposing non-negative constraints, basis matrix can be considered as latently embedded semantic topics and each column corresponds to one topic. Coefficient matrix can be accordingly considered as the relations between images and topics. For presentation convenience, we also term coefficient matrix as image-topic relation matrix. Formally, in modality m, the non-negative matrix factorization process can be formulated as

$$\min_{B_m, T_m} \|X_m - B_m T_m\|_{2,1} \text{ s.t. } B_m, T_m\big|_{m=1}^{2} \geq 0 \qquad (1)$$

where $B_m \in \mathbb{R}^{d_m \times K}$ and $T_m \in \mathbb{R}^{K \times N}$ are modality specific topic matrix and image-topic relation matrix, respectively, $||\cdot||_{2,1}$ is $l_{2,1}$ norm.[4] As illustrated in [53] and [54], $l_{2,1}$ norm is capable of resisting outliers of both images and texts by ensuring column-wise sparsity in the residual matrix $X_m - B_m T_m$. Besides, large reconstruction errors from noisy samples are not squared and do not dominate the objective value.

Since semantic topics are latently embedded in both visual and textual modalities, we impose an additional constraint to minimize the inconsistence between T_m and the shared image-topic relation matrix T. That is

$$\min_{T} \sum_{m=1}^{2} \alpha_m \|T_m - T\|_F^2 \text{ s.t.} \sum_{m=1}^{2} \alpha_m = 1, \alpha_m\big|_{m=1}^{2}, T \geq 0 \qquad (2)$$

where $\alpha_m|_{m=1}^{2}$ are weights which measure modality contributions. T actually reflects the relations between image and semantic topics, it is the main learning objective of image-topic discovery. By incorporating the above ideas, the overall image-topic relation discovery can be formulated as

$$\min_{T} \sum_{m=1}^{2} \Big(\|X_m - B_m T_m\|_{2,1} + \alpha_m \|T_m - T\|_F^2\Big)$$
$$\text{s.t. } B_m, T_m\big|_{m=1}^{2}, T, \alpha_m\big|_{m=1}^{2} \geq 0, \sum_{m=1}^{2} \alpha_m = 1. \qquad (3)$$

It is worth mentioning that the above equation is different from [55] which is developed for multiple non-negative matrix factorization.

1) Our formulation is specially designed for image-topic relation discovery. To the best of our knowledge, there is no similar work.
2) We impose $l_{2,1}$ norm instead of Frobenius norm on non-negative matrix factorization term. This design can well accommodate the noises and outliers involved in both visual contents and auxiliary texts.
3) The weights of non-negative matrix factorization terms can be automatically learned without any manual adjustment.

By solving (3), image-topic relation can be discovered. However, the overall framework involves $l_{2,1}$ norm which is nonsmooth and cannot be directly solved using a closed form. To bypass this problem, we transform it as the following alternative form:

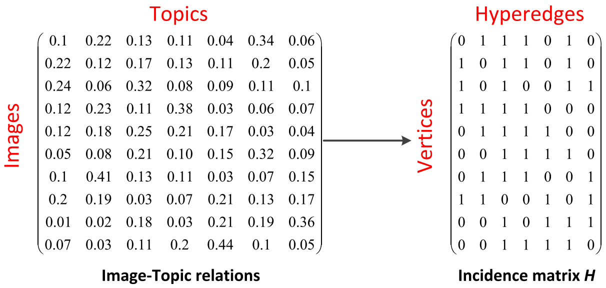
$$\min_{T} \sum_{m=1}^{2} \Big(\mathrm{Tr}\big((X_m - B_m T_m)\Lambda_m (X_m - B_m T_m)^{\mathrm{T}}\big)$$
$$+ \alpha_m \mathrm{Tr}\big((T_m - T)(T_m - T)^{\mathrm{T}}\big)\Big)$$
$$\text{s.t. } B_m, T_m\big|_{m=1}^{2}, T, \alpha_m\big|_{m=1}^{2} \geq 0, \sum_{m=1}^{2} \alpha_m = 1 \qquad (4)$$

where $\Lambda_m \in \mathbb{R}^{N \times N}$ is diagonal matrix, whose ith diagonal element is $(\Lambda_m)_{ii} = 1/2||y_i||_2$, y_i is the ith column of matrix $X_m - B_m T_m$. We adopt an iterative optimization to solve the problem (as shown in Algorithm 1). The following three steps are iterated until convergence.

[4] In this paper, for an example matrix $A = [a_1, \ldots, a_N]$, its $l_{2,1}$ norm $||A||_{2,1}$ is calculated as $\sum_{i=1}^{N} ||a_i||_2$.

Algorithm 1 Image-Topic Relation Discovery

```
Input:
    Feature matrices, $X_m|_{m=1}^2$. Number of semantic topics $K$.
Output:
    Image-topic relation matrix $T$.
Initialize $B_m, T_m|_{m=1}^2$, $T$, $\alpha_m|_{m=1}^2$.
while Eq.(4) Not Convergency do
    for $m = 1, 2$ do
        Compute $[y_1, \ldots, y_n] = X_m - B_m T_m$ and $\Lambda_m$.
        while Eq.(5) Not Convergency do
            Fixing $T$, $T_m$, and $\alpha_m$, calculate $B_m$ via Eq.(6).
            Fixing $T$, $B_m$, and $\alpha_m$, calculate $T_m$ via Eq.(7).
        end while
    end for
    Fixing other variables, calculate $T$ via Eq.(9).
    Fixing other variables, calculate $\alpha_m|_{m=1}^2$ via Eq.(13).
end while
```



$$\begin{pmatrix} 0.1 & 0.22 & 0.13 & 0.11 & 0.04 & 0.34 & 0.06 \\ 0.22 & 0.12 & 0.17 & 0.13 & 0.11 & 0.2 & 0.05 \\ 0.24 & 0.06 & 0.32 & 0.08 & 0.09 & 0.11 & 0.1 \\ 0.12 & 0.23 & 0.11 & 0.38 & 0.03 & 0.06 & 0.07 \\ 0.12 & 0.18 & 0.25 & 0.21 & 0.17 & 0.03 & 0.04 \\ 0.05 & 0.08 & 0.21 & 0.10 & 0.15 & 0.32 & 0.09 \\ 0.1 & 0.41 & 0.13 & 0.11 & 0.03 & 0.07 & 0.15 \\ 0.2 & 0.19 & 0.03 & 0.07 & 0.21 & 0.13 & 0.17 \\ 0.01 & 0.02 & 0.18 & 0.03 & 0.21 & 0.19 & 0.36 \\ 0.07 & 0.03 & 0.11 & 0.2 & 0.44 & 0.1 & 0.05 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Fig. 3. THG construction. Images and semantic topics are vertices and hyperedges, respectively.

Step 1: Fixing T, $\alpha_m|_{m=1}^2$, minimize the objective function in (4) with respect to $B_m|_{m=1}^2$ and $T_m|_{m=1}^2$

$$\begin{aligned} \min_{B_m, T_m|_{m=1}^2} \sum_{m=1}^{2} \Big(& \mathrm{Tr}\big((X_m - B_m T_m)\Lambda_m (X_m - B_m T_m)^{\mathrm{T}}\big) \\ & + \alpha_m \mathrm{Tr}\big((T_m - T)(T_m - T)^{\mathrm{T}}\big)\Big) \\ \text{s.t.} \ \ & B_m, T_m\big|_{m=1}^2 \geq 0. \end{aligned} \tag{5}$$

Theorem 1: Let $B_m, T_m|_{m=1}^2$ be defined as before. Then the updating rules for them can be expressed as

$$(B_m)_{dk} \leftarrow \frac{\big(X_m \Lambda_m T_m^{\mathrm{T}}\big)_{dk}}{\big(B_m T_m \Lambda_m T_m^{\mathrm{T}}\big)_{dk}} (B_m)_{dk} \tag{6}$$

$$(T_m)_{ki} \leftarrow \frac{\big(B_m^{\mathrm{T}} X_m \Lambda_m + \alpha_m T\big)_{ki}}{\big(B_m^{\mathrm{T}} B_m T_m \Lambda_m + \alpha_m T_m\big) ki} (T_m)_{ki}. \tag{7}$$

Proof: See Appendix A. ■

Step 2: Fixing $B_m, T_m|_{m=1}^2, \alpha_m|_{m=1}^2$, minimize the objective function in (4) with respect to T

$$\min_{T} \sum_{m=1}^{2} \alpha_m \|T_m - T\|_F^2 \ \text{s.t.} \ T \geq 0. \tag{8}$$

By setting the derivative of (8) with respect to T to 0, we get

$$\sum_{m=1}^{2} \alpha_m(-2T_m + 2T) = 0 \Rightarrow T = \sum_{m=1}^{2} \alpha_m T_m. \tag{9}$$

Step 3: Fixing $B_m, T_m|_{m=1}^2, T$, minimize the objective function in (4) with respect to $\alpha_m|_{m=1}^2$

$$\begin{aligned} \min_{\alpha_m|_{m=1}^2} \ & \sum_{m=1}^{2} \alpha_m \|T_m - T\|_F^2 \\ \text{s.t.} \ & \sum_{m=1}^{2} \alpha_m = 1, \alpha_m\big|_{m=1}^2 \geq 0. \end{aligned} \tag{10}$$

The above equation may lead to a trivial solution. To avoid it, we adopt similar trick used in [56] and [57]. We introduce a smooth factor $\xi > 1$ and set α_m to α_m^ξ, so that each modality can offer particular contribution to image-topic relation discovery. Meanwhile, with Lagrange multiplier μ, (10) can be transformed into the following equivalent optimization form to take into account the constraint of $\alpha_m|_{m=1}^2$:

$$\min_{\alpha_m|_{m=1}^2} \sum_{m=1}^{2} \alpha_m^\xi \|T_m - T\|_F^2 - \mu \left(\sum_{m=1}^{2} \alpha_m - 1 \right). \tag{11}$$

By setting the derivative of (11) with respect to α_m and μ to 0, we get

$$\begin{aligned} \frac{\partial \ (11)}{\alpha_m} &= \xi \alpha_m^{\xi - 1} \|T_m - T\|_F^2 - \mu = 0, m = 1, 2 \\ \frac{\partial \ (11)}{\mu} &= \sum_{m=1}^{2} \alpha_m - 1 = 0. \end{aligned} \tag{12}$$

Therefore, the solution of $\alpha_m|_{m=1}^2$ can be obtained

$$\alpha_m = \frac{\big(1/\|T_m - T\|_F^2\big)^{\frac{1}{\xi - 1}}}{\sum_{m=1}^{2} \big(1/\|T_m - T\|_F^2\big)^{\frac{1}{\xi - 1}}}. \tag{13}$$

Since $||T_m - T||_F^2 \geq 0$, we can naturally guarantee $\alpha_m \geq 0$.

C. Topic Hypergraph Construction

Semantic correlations of real images are complex and high-order. This paper is inspired to model them with a unified THG, where images and the semantic topics are considered as vertices and hyperedges, respectively. An illustration example is shown in Fig. 1 and a typical THG construction process is presented in Fig. 3. In this way, an image can represent several semantic topics, and several images jointly describe the same semantic topic. Besides, several images may be included in more than one topic. Hence, high-order semantic correlations are effectively modeled.

In this paper, THG $= (V, E, W)$ denotes THG, $V = \{v_i\}_{i=1}^N$ denotes the vertex set, $E = \{e_k\}_{k=1}^K$ denotes the hyperedge set, $W = \{w(e_k)\}_{k=1}^K$ denotes the weight set for hyperedges, $w(e_k)$ is the weight of hyperedge e_k. THG can be represented with a $N \times K$ incidence matrix H. For example, the element at ith row and kth column (the incidence value between vertex v_i and hyperedge e_k) is given as

$$H(v_i, e_k) = \begin{cases} 1 & T_{ki} \in \Gamma(v_i) \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where T_{ki} denotes the element at kth row and ith column of T. It characterizes the probability that an image describes the semantic topic. $\Gamma(v_i)$ is comprised of S largest elements in ith row of T. The degree of hyperedge is calculated as the number of images included. For e_k, its degree $\delta(e_k)$ is

$$\delta(e_k) = \sum_{i=1}^{N} H(v_i, e_k). \tag{15}$$

Weights of all hyperedges are set to 1, $w(e_k) = 1$, assuming that semantic topics are evenly distributed in the database. The degree of each vertex is defined as the sum of the weights of hyperedges that the vertex belongs to

$$d(v_i) = \sum_{k=1}^{K} w(e_k)H(v_i, e_k) = \sum_{k=1}^{K} H(v_i, e_k). \tag{16}$$

D. Hashing Learning

Hashing learning is performed on the constructed THG. Principally, images are more semantically similar if they are included to more identical hyperedges. They should be mapped into hashing codes with shorter Hamming distances. Moreover, we enforce the high-order semantic correlations of images to be preserved in the binary hashing codes. To achieve this goal, we explicitly minimize the distance between the incidence matrix and hashing codes. By considering these two parts, the objective function can be formulated as[5]

$$\begin{aligned}
\min_{Z}\ & \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} H'(v_i, e_k)\sum_{k'=1}^{K}\left(\frac{z_{k'i}}{\sqrt{d(v_i)}} - \frac{z_{k'j}}{\sqrt{d(v_j)}}\right)^2 \\
& + \mu\left(\sum_{k=1}^{K}\sum_{i=1}^{N}(z_{ki} - H(v_i, e_k))^2\right) \\
\text{s.t. } & H'(e_k, v_i) = \frac{w(e_k)H(v_i, e_k)H(v_j, e_k)}{\delta(e_k)}, z_i \in \{-1, 1\}^K
\end{aligned} \tag{17}$$

where $\mu, \beta, \gamma > 0$ are factors that adjust the balance between regularization terms. The first term is hypergraph Laplacian constraint, while the second term explicitly preserves the extracted high-order semantic relations.

The first term in (17) can be transformed as

$$\begin{aligned}
&\frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{N}\frac{w(e_k)H(v_i, e_k)H(v_j, e_k)}{\delta(e_k)}\sum_{k'=1}^{K}\left(\frac{z_{k'i}}{\sqrt{d(v_i)}} - \frac{z_{k'j}}{\sqrt{d(v_j)}}\right)^2 \\
&= \sum_{i=1}^{N}\left(\sum_{k=1}^{K} z_{k'i}^2\right)\sum_{k=1}^{K}\frac{w(e_k)H(v_i, e_k)}{d(v_i)}\sum_{j=1}^{N}\frac{H(v_j, e_k)}{\delta(e_k)} \\
&\quad - \sum_{k=1}^{K}\sum_{i,j=1}^{N}\frac{w(e_k)H(v_j, e_k)H(v_i, e_k)\left(\sum_{k'=1}^{K} z_{k'i}z_{k'j}\right)}{\delta(e_k)\sqrt{d(v_j)d(v_i)}}.
\end{aligned} \tag{18}$$

From (15) and (16), we can obtain that

$$\sum_{j=1}^{N}\frac{H(v_j, e_k)}{\delta(e_k)} = 1 \quad \sum_{k=1}^{K}\frac{w(e_k)H(v_i, e_k)}{d(v_i)} = 1. \tag{19}$$

Then, the first term in (17) can be represented as

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\sum_{k'=1}^{K} z_{k'i}^2\right) - \sum_{k=1}^{K}\sum_{i,j=1}^{N} \\
&\quad \times \frac{w(e_k)H(v_j, e_k)H(v_i, e_k)\left(\sum_{k'=1}^{K} z_{k'i}z_{k'j}\right)}{\delta(e_k)\sqrt{d(v_j)d(v_i)}} \\
&= \sum_{k'=1}^{K}\left(\sum_{i=1}^{N} z_{k'i}^2 - \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{N}\frac{w(e_k)H(v_j, e_k)H(v_i, e_k)z_{k'i}z_{k'j}}{\delta(e_k)\sqrt{d(v_j)d(v_i)}}\right) \\
&= \sum_{k'=1}^{K}\left(z_{k'\cdot}\Delta z_{k'\cdot}^{\mathrm{T}}\right) = \mathrm{Tr}\left(Z\Delta Z^{\mathrm{T}}\right)
\end{aligned} \tag{20}$$

where $z_{k'\cdot}$ denotes k'th row of matrix Z, $\Delta \in \mathbb{R}^{N\times N}$ is Laplacian matrix of THG, it is calculated as

$$\Delta = I - D_v^{-1/2}HD_wD_e^{-1}H^TD_v^{-1/2} \tag{21}$$

where D_v, D_e, and D_w are the diagonal matrices of the vertex degrees, edge degrees, and hyperedge weights, respectively. Equation (17) can be represented as a compact form

$$\min_{Z} \mathrm{Tr}\left(Z\Delta Z^{\mathrm{T}}\right) + \mu\left\|Z - H^T\right\|_F^2 \text{ s.t. } Z \in \{-1, 1\}^{K\times N}. \tag{22}$$

Considering the limited computational resources of mobile ends, we leverage linear regression to learn hashing function

$$\min_{P} \left\|Z - P^{\mathrm{T}}X_1\right\|_F^2 + \gamma||P||_F^2 \tag{23}$$

where P is the projection matrix which maps raw visual feature into hashing codes. With this design, the online hashing process can be efficiently completed with linear projection operations.

We integrate hashing code and function learning into a unified framework. The objective function becomes

$$\begin{aligned}
\min_{Z,P}\ & \mathrm{Tr}\left(Z\Delta Z^{\mathrm{T}}\right) + \mu\left\|Z - H^{\mathrm{T}}\right\|_F^2 \\
& + \beta\left(\left\|Z - P^{\mathrm{T}}X_1\right\|_F^2 + \gamma||P||_F^2\right) \\
\text{s.t. } & Z \in \{-1, 1\}^{K\times N}.
\end{aligned} \tag{24}$$

Solving the above problem is NP-hard due to discrete constraint $Z \in \{-1, 1\}^{K\times N}$. To make it computationally tractable, we relax this constraint and obtain continuous solution. The final discrete solution can be effectively calculated by mean thresholding. As indicated by [9] and [36], mean thresholding can balance the partition of database and provide maximum information. It should be noted that, to guarantee the explicit semantic correlation at the same scale, we also relax incidence matrix H^{T} to T accordingly. Therefore, the relaxed objective function is transformed as

$$\begin{aligned}
\min_{Z,P}\ & \mathrm{Tr}\left(Z\Delta Z^{\mathrm{T}}\right) + \mu||Z - T||_F^2 \\
& + \beta\left(\left\|Z - P^{\mathrm{T}}X_1\right\|_F^2 + \gamma||P||_F^2\right).
\end{aligned} \tag{25}$$

Note that the above equation differs from the semi-supervised setting of prior work flexible manifold embedding in [58] on different intrinsic structure and meaning.

[5] We substitute 0 in incidence H with -1 in objective function.

Algorithm 2 Iterative Computation for Z

```
Input:
    Image-topic relation matrix $T$.
Output:
    Hashing codes of images $Z = Z^t$.
Initialize $Z^1 = T, t = 1$;
repeat
    $t = t + 1$;
    Update $Z$ via: $Z^t = \frac{1}{1+\lambda} Z^{t-1}(I - \Delta') + T\frac{\lambda}{1+\lambda}$;
until Convergency
return $Z^t$.
```

1) The Laplacian matrix Δ in (25) is calculated from THG which is specially designed in this paper to model high-order semantic correlations of images.
2) The above equation is built within unsupervised learning framework. T is obtained in image-topic relation discovery by jointly modeling images and auxiliary texts. The whole process requires no manual semantic labels.

We set the derivative (25) with respect to P to 0, and have

$$2X_1X_1^{\mathrm{T}}P - 2ZX_1^{\mathrm{T}} + 2\gamma P = 0$$
$$\Rightarrow P = \left(X_1X_1^{\mathrm{T}} + \gamma I_K\right)^{-1}X_1Z^{\mathrm{T}}. \quad (26)$$

Let $M = (X_1X_1^{\mathrm{T}} + \gamma I_K)^{-1}$, then $P = MX_1Z^{\mathrm{T}}$. we obtain

$$\begin{aligned}
&\left\|Z - P^{\mathrm{T}}X_1\right\|_F^2 + \gamma||P||_F^2\\
&\quad = \mathrm{Tr}\big(Z - ZX_1^{\mathrm{T}}MX_1\big)\big(Z - ZX_1^{\mathrm{T}}MX_1\big)^{\mathrm{T}}\\
&\qquad + \gamma\mathrm{Tr}\big(ZX_1^{\mathrm{T}}M^{\mathrm{T}}MX_1Z^{\mathrm{T}}\big)\\
&\quad = \mathrm{Tr}\big(Z\big(I_N - 2X_1^{\mathrm{T}}MX_1 + X_1^{\mathrm{T}}M\big(X_1X_1^{\mathrm{T}} + \gamma I_K\big)MX_1\big)Z^{\mathrm{T}}\big)\\
&\quad = \mathrm{Tr}\Big(Z\Big(I_N - 2X_1^{\mathrm{T}}MX_1 + X_1^{\mathrm{T}}MM^{-1}MX_1\Big)Z^{\mathrm{T}}\Big)\\
&\quad = \mathrm{Tr}\big(Z\big(I_N - X_1^{\mathrm{T}}MX_1\big)Z^{\mathrm{T}}\big).
\end{aligned}$$

The optimization formula in (25) is derived as

$$\min_Z \ \mathrm{Tr}\big(Z\Delta Z^{\mathrm{T}}\big) + \mu||Z - T||_F^2 + \beta\mathrm{Tr}\big(Z\big(I_N - X_1^{\mathrm{T}}MX_1\big)Z^{\mathrm{T}}\big)$$
$$\Leftrightarrow \min_Z \mathrm{Tr}\big(Z\Delta'Z^{\mathrm{T}}\big) + \mu||Z - T||_F^2 \quad (27)$$

where $\Delta' = \Delta + \beta(I_N - X_1^{\mathrm{T}}MX_1)$. By calculating the derivative of (27) with respect to Z and set it to 0, we derive

$$Z = T\left(I_N + \frac{1}{\mu}\Delta'\right)^{-1}. \quad (28)$$

Similar to the method developed in [59], the above equation can be effectively calculated via an iterative process. The detailed steps are illustrated in Algorithm 2. The convergency proof is presented in Appendix B.

By substituting (28) into (26), we obtain projection matrix P. Following the rules of mean thresholding, we first calculate the mean projected vector of database images $b = (P^{\mathrm{T}}X_1 1_N/N)$, and then construct hashing functions as

$$G(x) = \frac{\mathtt{sgn}\big(P^{\mathrm{T}}x - b\big) + 1}{2}. \quad (29)$$

The behind meaning of the above equation is: the projected feature dimension that is larger than the specified threshold is remapped to 1 via hashing function, and 0 vice versa.

Algorithm 3 THH-Based CBMIR

```
Input:
    Query image, $q$.
    Database images, $\{I_i\}_{i=1}^N$.
    Non-negative hyper-parameters, $\mu, \beta, \gamma$.
    Number of topics that images describe, $S$.
    Hashing code length, $K$.
Output:
    Hashing codes of database images, $Z$.
    Hashing functions, $G$.
    Image retrieval results for image query $q$.
    Offline Learning
 1: Extract features $X_m|_{m=1}^2$ of database images;
 2: Compute image-topic relation via Algorithm 1;
 3: Construct THG as illustrated in Section IV-C;
 4: Compute Laplacian matrix of THG via Eq.(21);
 5: Compute $M = (X_1X_1^{\mathrm{T}} + \gamma I_K)^{-1}$;
 6: Compute hashing codes of database images $Z$ via Eq.(27) and
    Algorithm 2;
 7: Construct hashing functions $G$ via Eq.(29);
    Online Retrieval
 8: Extract visual feature of query image;
 9: Project query into hashing codes via Eq.(27);
10: Perform retrieval is Hamming space and return results.
```

E. Summary and Computation Complexity Analysis

The key steps of THH-based CBMIR are described in Algorithm 3. The computation cost consists of two major parts: 1) offline training and 2) online CBMIR. It can be easily derived that the computation cost of image-topic relation discovery is $O(\mathrm{Iter}_1 \cdot \mathrm{Iter}_2 \cdot (d_1 + d_2) \cdot N \cdot K)$, where Iter_1 and Iter_2 denote the number of iterations for steps 3–11 and steps 7–8 in Algorithm 1, respectively. The process of THG construction consumes $O(N \cdot S)$. The complexity of calculating the inverse of matrix in step 5 is $O(K^3)$. Solving (27) has computation complexity of $O(\mathrm{Iter}_3 \cdot N \cdot K)$, where Iter_3 is the number of iterations for computing Z. The process of hashing code generation for database images costs time complexity $O(N)$. In a sum, the whole offline training consumes time complexity $O(N)$, which is linear to database image size. In online retrieval, generating hashing codes for a query can be completed in $O(d_1 \cdot K + K)$, which is quite efficient. The search process can be efficiently completed with simple bit operations.

V. Experimental Configuration

A. Experimental Datasets

In this paper, we empirically evaluate the performance of THH on two publicly available multimodal image datasets: Wiki [24] and NUS-WIDE [60]. To the best of our knowledge, there is still no publicly available full-labeled multimodal mobile image datasets. In experiments, we use both datasets to model the real application scenario of CBMIR, where query is only visual image and database images are usually associated with noisy but informative texts.

1) Wiki[6] consists of 2866 multimedia documents in ten semantic categories. Visual and textual contents are represented by 1000 dimensional bag-of-visual-words and 2000 dimensional bag-of-words, respectively. Images are considered to be relevant only if they belong to the same

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

category. Five percent images are used as query set and the remaining images are used as training and database set.

2) NUS-WIDE[7] is comprised of 269 648 images which are labeled into 81 concepts. We preserve ten most common concepts and the corresponding 186 643 pairs. Images are represented by 500 dimensional bag-of-visual-words. Textual features are 1000 dimensional binary bag-of-words. Images are considered to be relevant if they share at least one concept. One percent images are used as query set, 3% images are used as training set, and the remaining images are used as database set.

On both datasets, the hashing codes learned on training set are all discarded after hashing function learning. The constructed hashing functions are leveraged to generate hashing codes for both query and database images.

B. Evaluation Metrics

In our experimental study, mean average precision (mAP) is adopted as the evaluation metric for effectiveness. The metric has been widely used in [22]. For a given query, average precision (AP) is calculated as

$$\mathrm{AP} = \frac{1}{\mathrm{NR}} \sum_{r=1}^{R} \mathrm{pre}(r)\mathrm{rel}(r) \tag{30}$$

where R is the total number of retrieved images, NR is the number of relevant images in retrieved set, pre(r) denotes the precision of top r retrieval images, which is defined as the ration between the number of the relevant images and the number of retrieved images r, and rel(r) is indicator function which equals to 1 if the rth image is relevant to query, and 0 vice versa. mAP is defined as the average of the AP of all queries. Larger mAP means the retrieval performance is better. In experiments, we set R as 100 to collect experimental results. Furthermore, precision–scope curve is also reported to reflect the retrieval performance variations with respect to the number of retrieved images.

C. Compared Approaches

THH is unsupervised. Hence, we compare it with several state-of-the-art unsupervised uni- and cross-modal approaches. UMH approaches used for comparison include: shift-invariant kernels LSH (SKLSH) [15], SPH [8], PCAH [33], LSI [34], AGH [37], SRH [35], STH [36], and ITQ [16]. CMH approaches used for evaluation include[8] the following.

1) *CVH [20]:* It extends SPH to learn hashing functions by jointly minimizing Hamming distances of similar samples and maximizing that of dissimilar samples.

2) *CHMIS [21]:* It integrates discriminative information from several heterogeneous modalities into the binary hashing codes with proper weights. For comparison fairness, text input is removed and only visual query is preserved in CHMIS. In this case, CHMIS can also be considered as CMH.

3) *IMH [22]:* It formulates hashing learning in a framework where intrasimilarity of each individual modality and intercorrelations between different modalities are both preserved in hashing codes.

4) *CMFH [62]:* It learns a latent semantic subspace shared by multiple modalities by collective matrix factorization. In CMFH, both visual and text features are mapped into a unified hashing codes.

Note that, CVH, CHMIS, IMH, and CMFH generate hashing codes for both query image and text. Since we only test the performance of CBMIR in experiments, the hashing codes of text are removed. In this case, the retrieval process of CBMIR in all compared approaches is performed in visual Hamming space. Parameters of all compared approaches are adjusted to maximize the performance according to the relevant literature.

D. Implementation Details

The hyperparameters of THH are tuned by standard parallel grid-search on a subset of training data. S is used in (14) to control the number of topics that images describe. The best performance of THH is achieved when $S = 4$ on both datasets. In (17), there are three parameters: μ, β, and γ, which adjust balance between regularization terms. These parameters are chosen from $\{10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$ in this paper. In particular, the best performance of THH is achieved when $\{\mu = 10^4, \beta = 10^{-2}, \gamma = 10^{-4}\}$, $\{\mu = 10^2, \beta = 10^2, \gamma = 10^4\}$ on Wiki and NUS-WIDE, respectively. In experiments, hashing code length L on all datasets is varied in the range of $\{16, 32, 64, 128\}$ to observe the performance. The retrieval scope on Wiki is set from 100 to 1000 with step size 100, that on NUS-WIDE is set from 500 to 5000 with step size 500.

In step 1 of Algorithm 1, the initial values of $B_m|_{m=1}^2$ and $T_m|_{m=1}^2$ are obtained by non-negative matrix factorization on feature matrix $X_m|_{m=1}^2$, T is calculated as the mean of T_1 and T_2. All the experiments are conducted on a computer with Intel(R) Xeon(R) CPU E5-2620 2.0 GHz and 32 GB RAM.

VI. Experimental Results and Discussions

In this section, we first present the performance comparison results. Then, we give comprehensive analysis and discussion about how various factors influence the performance of THH. In particular, we investigate the effects of THG learning, text assistance and $l_{2,1}$ norm on the overall system performance. Finally, we study the performance variations of THH with involved parameters.

A. Performance Comparison

Tables III and IV present main mAP results of THH and all compared approaches on Wiki and NUS-WIDE when different code length is set. Their corresponding precision–scope curves are demonstrated in Figs. 4 and 5, respectively. According to

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

[8] For SKLSH, SPH, PCAH, LSI, STH, AGH, and ITQ, implementation codes of them are provided by [36]. For SRH, we carefully implement the code according to the relevant literature. For CVH, we kindly use the implementation code provided by [61]. For CHMIS, IMH, and CMFH, we directly download implementation codes from author website.

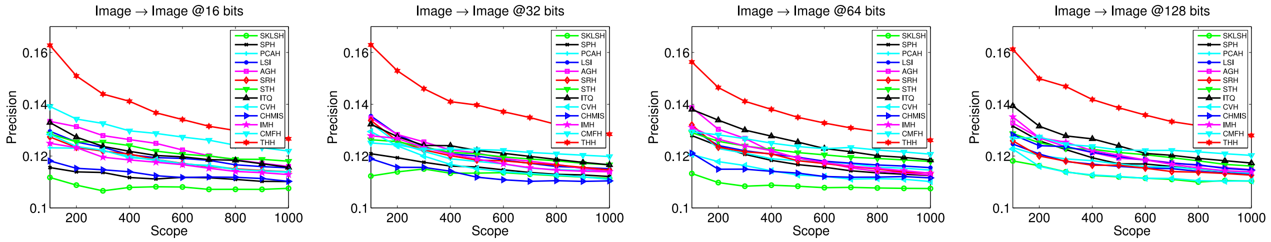


Fig. 4. Precision–scope curves on Wiki varying code length.

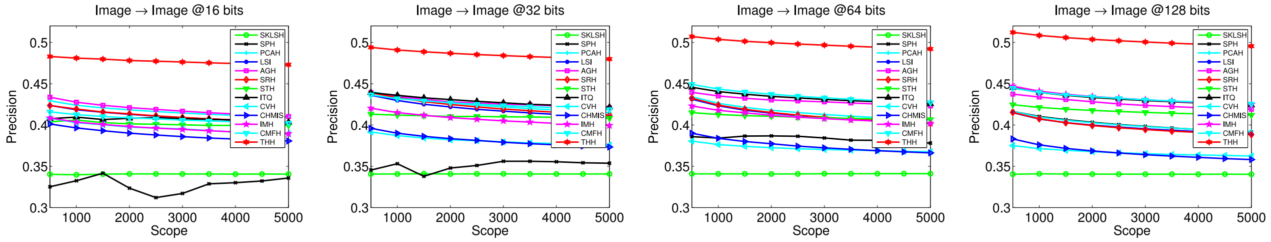


Fig. 5. Precision–scope curves on NUS-WIDE varying code length.

TABLE III
MAP OF ALL APPROACHES ON WIKI

Methods	Wiki			
	16 bits	32 bits	64 bits	128 bits
SKLSH	0.1433	0.1530	0.1501	0.1681
SPH	0.1535	0.1677	0.1742	0.1727
PCAH	0.1641	0.1771	0.1752	0.1702
LSI	0.1683	0.1777	0.1715	0.1662
AGH	0.1754	0.1645	0.1748	0.1802
SRH	0.1696	0.1786	0.1796	0.1747
STH	0.1740	0.1706	0.1718	0.1745
ITQ	0.1742	0.1753	0.1871	0.1900
CVH	0.1730	0.1791	0.1689	0.1668
CHMIS	0.1539	0.1732	0.1648	0.1802
IMH	0.1647	0.1605	0.1676	0.1766
CMFH	0.1719	0.1661	0.1679	0.1721
THH	**0.2033**	**0.2021**	**0.2054**	**0.2051**

TABLE IV
MAP OF ALL APPROACHES ON NUS-WIDE

Methods	NUS-WIDE			
	16 bits	32 bits	64 bits	128 bits
SKLSH	0.3685	0.3695	0.3680	0.3673
SPH	0.3430	0.3910	0.4415	0.4583
PCAH	0.4653	0.4817	0.4789	0.4632
LSI	0.4645	0.4803	0.4746	0.4600
AGH	0.4646	0.4681	0.4711	0.4731
SRH	0.4614	0.4824	0.4758	0.4603
STH	0.4120	0.4368	0.4443	0.4606
ITQ	0.4482	0.4667	0.4860	0.4859
CVH	0.4447	0.4300	0.4233	0.4149
CHMIS	0.4419	0.4347	0.4302	0.4265
IMH	0.4475	0.4618	0.4634	0.4879
CMFH	0.4703	0.4883	0.4942	0.4882
THH	**0.5058**	**0.5218**	**0.5375**	**0.5447**

the presented results, we can clearly find that THH can consistently achieve superior retrieval performance compared with competitors. The largest performance gap between THH and the second best performance is more than 5% on NUS-WIDE. It demonstrates that, with the assistance of text, THH can enrich the semantics of hashing codes and improve retrieval performance. Even with 16 bits, THH can obtain better performance than the one obtained by the compared approach on 128 bits. This desirable advantage shows that THH can well support the CBMIR scenario, where network transmission bandwidth is limited.

Besides, it is interesting to find that CMH approaches even perform worse than UMH methods (for example, ITQ and AGH) in several cases. This experimental phenomenon is not consistent with the conclusion obtained in [22] that CMH methods perform better than UMH methods on task of uni-modal retrieval. We think the reason is that CMH aims to achieve fast retrieval across heterogeneous modalities. Therefore, seeking the shared space of heterogenous modalities is the main aim. In this way, the discovered common semantic space of heterogeneous modalities in CMH can principally preserve semantic correlations of different modalities. But, it may even lose the valuable semantics besides the common part in original visual features. CMH may not be the best suited for CBMIR. This observation also motivates us to design THH to effectively leverage the auxiliary text to assist image hashing.

We also investigate the efficiency of online image retrieval. As indicated in Algorithm 3, the online retrieval process is comprised of three subsequent steps 8–10. Since steps 8 and 10 are identical for all compared approaches, we only test the efficiency of step 9. In particular, we compare the hashing code generation time of all query images. Table V presents the main experimental results. From it, we can easily find that THH consumes the least time on both datasets. This is because THH is designed to generate the hashing codes via linear projection, while most of the compared approach require to include additional procedures to improve performance. The results demonstrate that THH can achieve higher retrieval accuracy with simple online computation operations. This desirable advantage of THH can well support its application for efficient CBMIR, where computational resources in mobile end are quite limited and scarce, and

TABLE V

HASHING CODE GENERATION TIME (S) WHEN HASHING CODE LENGTH IS FIXED TO 128

	SKLSH	SPH	PCAH	LSI	AGH	SRH	STH	ITQ	CVH	CHMIS	IMH	CMFH	THH
Wiki	0.002	0.056	0.003	0.002	0.005	0.039	0.569	**0.001**	0.002	0.041	0.003	**0.001**	**0.001**
NUS-WIDE	0.018	0.394	0.015	0.016	0.053	0.646	5.641	**0.008**	0.01	0.2381	0.02	**0.008**	**0.008**

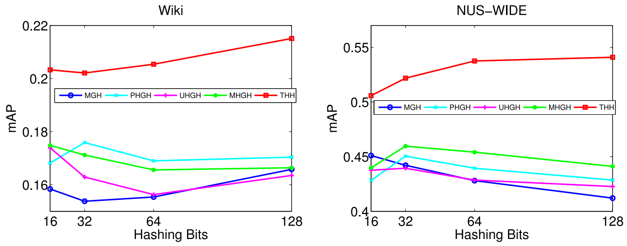


Fig. 6. Hashing performance based on state-of-the-art hypergraphs.

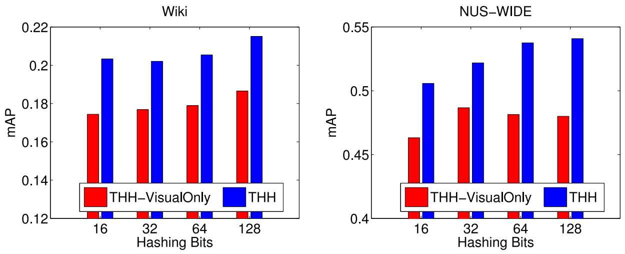


Fig. 7. Effects of text assistance on hashing performance.

online retrieval efficiency plays a great impact on the user's experience.

B. Effects of Topic Hypergraph Learning

THG learning is performed in this paper to model the high-order semantic relations that actually exist among images. This section conducts experiments to evaluate the effects of topic hyeprgraph learning on hashing. To this end, we compare the hashing performance achieved on state-of-the-art hypergraphs. In particular, the following hypergraph variants are compared.

1) *Multimodal Graph (MG) [63]:* MG constructs simple graph in each modality and then combines them into a unified one to model the simple semantic relations among images. The optimal combination weights are calculated by brute force search from 0–1 with step-size 0.1. The aim of comparing THH with MG is to validate the effects of high-order relation modeling on hashing performance.
2) *Unified Hypergraph (UHG) [64]:* UHG constructs hypergraph by integrating hyperedges in multimodal hypergraphs (MHGs) into a UHG.
3) *PHG [48]:* PHG assigns each vertex to hyperedge in probabilistic way. Multimodal features are combined to describe the affinity relations among vertices within each hyperedge.
4) *MHG [65]:* MHG learns proper weights for multiple hypergraphs, and combines them into a unified one. It comprehensively considers high-order relations captured multiple modalities. The optimal combination weights are calculated with the same way as MG.

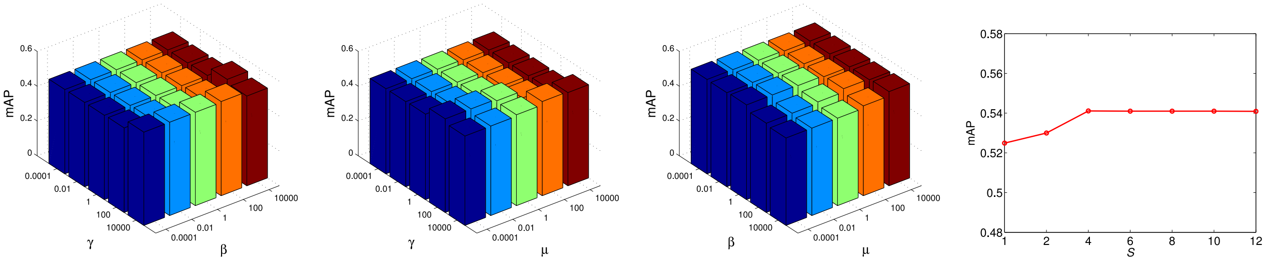
The optimal hashing codes based on the compared hypergrpah approaches are obtained by substituting the THG Laplacian with the corresponding one. For illustration convenience, the above hypergraph-based hashing approaches are denoted as MGH, UHGH, PHGH, and MHGH, respectively. Fig. 6 presents the main compared results. We easily find that THH can achieve superior performance than the competitors. In addition, we can observe that MGH obtains the worst performance. This significant performance gap clearly validates the effects of high-order semantic relation discovery on enhancing the semantic representation capability of hashing codes. Besides, among hypergraph-based hashing techniques, THH can still achieve the best performance. The reason is that THH can discover latent semantic topics. And in the process of hashing learning, it can explicitly preserve the modeled high-order semantic correlation in hashing codes. This advantage of THH can effectively mitigate the semantic shortage of hypergraph-based hashing techniques, which are directly built on low-level features.

C. Effects of Text Assistance

Auxiliary texts are exploited in this paper to detect semantic topics and construct THG. The main aim of this experiment is to investigate the effects of text assistance. To achieve this goal, we compare the performance of THH with the compared method which ignores semantics in texts and only considers visual information. For illustration convenience, we denote this compared approach as THH-VisualOnly. In implementation, THH-VisualOnly constructs THG by considering only visual information and learns the hashing codes as THH. Fig. 7 presents the detailed empirical experimental results. The key observations we gain are twofold: first, THH can achieve better retrieval performance of CBMIR with the assistance of texts. The potential reason is that, with text assistance, more valuable semantics can be involved into the detected semantic topics, and the constructed THG can better characterize the high-order semantic correlations of images. Hence, the generated hashing codes have better discriminative capability. Second, performance gap is varied on different datasets and hashing code lengths. The variations of performance gap are mainly caused by the different effectiveness of texts on assisting hashing.

D. Effects of $l_{2,1}$ Norm

We also evaluate the effectiveness of $l_{2,1}$ norm on learning hashing codes. We compare the performance of THH with the competitor which adopts Frobenius norm to discover image-topic relation [Frobenius norm is imposed on residual matrix $X_m - B_m T_m$ in (3) instead of $l_{2,1}$ norm]. For presentation



(a)

(b)

(c)

(d)

Fig. 8. THH performance variations with parameters. (a) μ is fixed to 10^2. (b) β is fixed to 10^2. (c) γ is fixed to 10^4. (d) Variations with S.

TABLE VI
EFFECTS OF $l_{2,1}$ NORM ON LEARNING HASHING CODES

Wiki	16 bits	32 bits	64 bits	128 bits
THH-F	0.1799	0.1859	0.1901	0.1901
THH	**0.2033**	**0.2021**	**0.2054**	**0.2051**

NUS-WIDE	16 bits	32 bits	64 bits	128 bits
THH-F	0.4899	0.5130	0.5236	0.5277
THH	**0.5058**	**0.5218**	**0.5375**	**0.5447**

convenience, we denote the compared approach THH-F. Table VI presents the main comparison results on two datasets. From it, we clearly find that THH can consistently achieve better performance than THH-F. The results demonstrate that $l_{2,1}$ norm performs better than Frobenius norm on enhancing the robustness of hashing codes.

E. Parameter Sensitivity

In this section, we conduct empirical experiments to study the performance variations with involved parameters in THH. More specifically, we observe the performance variations of THH with μ, β, γ, and S. μ, β, γ are used in (17) to play tradeoff between regularization terms and empirical loss, S is used in (14) to control the number of topics that images describe. Due to the limited space, we only report the results on NUS-WIDE when hashing code length is 128. Similar results can be found on other code lengths and Wiki. We test the results when μ, β, γ are varied from $[10^{-4}, 10^{-2}, 1, 10^2, 10^4]$, S is varied from $\{1, 2, 4, 6, 8, 10, 12\}$. Since μ, β, γ are used in the same equation, we observe performance variations with two of them by fixing the other parameter. Experimental results are presented in Fig. 8. From this figure, we can clearly find that the performance is relatively stable to a wide range of μ, β, γ variations. Besides, we can observe that the performance first increases with S and becomes stable after a certain point.

VII. CONCLUSION

CBMIR is a practical retrieval paradigm to support convenient mobile image retrieval. Hashing can be applied as an effective technique to facilitate large-scale CBMIR, due to its efficient transmission, low storage cost and fast retrieval response. However, most existing hashing methods are designed based on pure visual statistical information without considering the informative text, which is usually associated with Web images. Although CMH can potentially leverage texts, it still fails to fully make use of the texts. This paper proposes a novel unsupervised hashing to specially leverage auxiliary texts to imporve the effectiveness of hashing in visual space. We learn hashing codes and functions within a unified THH framework, which models high-order semantic correlations of images and preserves them in the hashing codes via unsupervised learning. This design has desirable advantages of convenient query input and high quality feature compression. Thus, it can well support practical applications of mobile image retrieval. The results gained from experiments on standard image datasets demonstrate the promising effectiveness of the proposed approach.

In the future, we plan to construct large-scale multimodal mobile image datasets and evaluate the performance of different approaches. Besides, it would also be interesting to apply the proposed method to other practical applications, which have similar characteristics with mobile image retrieval.

APPENDIX A
PROOF OF THEOREM 1

With Lagrange multiplier, (5) can be transformed as the following unconstraint optimization problem:

$$\min_{T} \sum_{m=1}^{2}\Big(\mathrm{Tr}\big((X_m - B_mT_m)\Lambda_m(X_m - B_mT_m)^{\mathrm{T}}\big) + \alpha_m Tr\big((T_m - T)(T_m - T)^{\mathrm{T}}\big) + \mathrm{Tr}\big(\Phi_m B_m^{\mathrm{T}}\big) + \mathrm{Tr}\big(\Psi_m T_m^{\mathrm{T}}\big)\Big)$$

where $\Phi_m = [\phi_{dk}], \Psi_m = [\psi_{ki}], d = 1, \ldots, d_m, k = 1, \ldots, K, i = 1, \ldots, N$ $\phi_{dk} > 0$, $\psi_{ki} > 0$ control the constraint of $(B_m)dk > 0$ and $(T_m)ki > 0$, respectively. By setting the derivative of (5) with respect to B_m, T_m to 0, we get

$$-2X_m\Lambda_mT_m^{\mathrm{T}} + 2B_mT_m\Lambda_mT_m^{\mathrm{T}} + \Phi = 0$$
$$-2B_m^{\mathrm{T}}X_m\Lambda_m + 2B_m^{\mathrm{T}}B_mT_m\Lambda_m + 2\alpha_m(T_m - T) + \Psi = 0.$$

Using the KKT conditions [66], $\phi_{dk}b_{dk} = 0$, $\psi_{ki}t_{ki} = 0$, we can derive the following equations:

$$\big(-X_m\Lambda_mT_m^{\mathrm{T}} + B_mT_m\Lambda_mT_m^{\mathrm{T}}\big)_{dk}B_{dk} = 0$$
$$\big(-B_m^{\mathrm{T}}X_m\Lambda_m + 2B_m^{\mathrm{T}}B_mT_m\Lambda_m + \alpha_m(T_m - T)\big)_{ki}T_{ki} = 0.$$

Hence, according to the standard procedure of non-negative matrix factorization [52], we can obtain the updating rules for $(B_m)_{dk}$ and $(T_m)_{ki}$

$$(B_m)_{dk} \leftarrow \frac{\left(X_m \Lambda_m T_m^{\mathrm{T}}\right)_{dk}}{\left(B_m T_m \Lambda_m T_m^{\mathrm{T}}\right)_{dk}} (B_m)_{dk}$$

$$(T_m)_{ki} \leftarrow \frac{\left(B_m^{\mathrm{T}} X_m \Lambda_m + \alpha_m T\right)_{ki}}{\left(B_m^{\mathrm{T}} B_m T_m \Lambda_m + \alpha_m T_m\right) ki} (T_m)_{ki}.$$

Appendix B
Proof of the Convergence of Algorithm 2

Z at the t_{th} iteration can be calculated as

$$Z^t = \left(\frac{\mu}{1+\mu}\right) \sum_{i=0}^{t-1} T \left(\frac{1}{1+\mu}(I - \Delta')\right)^i + T\left(\frac{1}{1+\mu}(I - \Delta')\right)^t.$$

Since the eigenvalues of $I - \Delta'$ are $[1, -1]$, we obtain that

$$\lim_{t \to \infty} \sum_{i=0}^{t-1} T \left(\frac{1}{1+\mu}(I - \Delta')\right)^i = \left(I - \frac{1}{1+\mu}(I - \Delta')\right)^{-1} = \frac{1+\mu}{\mu}\left(I + \frac{1}{\mu}\Delta'\right)^{-1}, \lim_{t \to \infty} T\left(\frac{1}{1+\mu}(I - \Delta')\right)^t = 0.$$

Therefore, we can derive that

$$Z = \lim_{t \to \infty} Z^t = \left(\frac{\mu}{1+\mu}\right) \lim_{t \to \infty} \sum_{i=0}^{t-1} T\left(\frac{1}{1+\mu}(I - \Delta')\right)^i + \lim_{t \to \infty} T\left(\frac{1}{1+\mu}(I - \Delta')\right)^t = T\left(I + \frac{1}{\mu}\Delta'\right)^{-1}.$$

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive and helpful suggestions.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, p. 5, 2008.

[2] R. Ji *et al.*, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2012.

[3] J. He *et al.*, "Mobile product search with bag of hash bits and boundary reranking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2012, pp. 3005–3012.

[4] L.-Y. Duan, R. Ji, Z. Chen, T. Huang, and W. Gao, "Towards mobile document image retrieval for digital library," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 346–359, Feb. 2014.

[5] B. Girod *et al.*, "Mobile visual search," *IEEE Signal Process. Mag.*, vol. 28, no. 4, pp. 61–76, Jul. 2011.

[6] J. Wang, H. T. Shen, J. Song, and J. Ji, "Hashing for similarity search: A survey," *CoRR*, vol. abs/1408.2927, Aug. 2014.

[7] L. Gao *et al.*, "Learning in high-dimensional multimedia data: The state of the art," *Multimedia Syst.*, pp. 1–11, Oct. 2015, doi: 10.1007/s00530-015-0494-1.

[8] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2008, pp. 1753–1760.

[9] J. Song, Y. Yang, X. Li, Z. Huang, and Y. Yang, "Robust hashing with local models for approximate similarity search," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1225–1236, Jul. 2014.

[10] Y. Zhen, Y. Gao, D.-Y. Yeung, H. Zha, and X. Li, "Spectral multimodal hashing and its application to multimedia retrieval," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 27–38, Jan. 2016.

[11] R. Ye and X. Li, "Compact structure hashing via sparse and similarity preserving embedding," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 718–729, Mar. 2016.

[12] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, "Large-scale unsupervised hashing with shared structure learning," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1811–1822, Sep. 2015.

[13] L. Chen, D. Xu, I. W.-H. Tsang, and X. Li, "Spectral embedded hashing for scalable image retrieval," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1180–1190, Jul. 2014.

[14] Z. Jin, C. Li, Y. Lin, and D. Cai, "Density sensitive hashing," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1362–1371, Aug. 2014.

[15] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2009, pp. 1509–1517.

[16] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[17] J. Song *et al.*, "A distance-computation-free search scheme for binary code databases," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 484–495, Mar. 2016.

[18] Y. Mu, J. Shen, and S. Yan, "Weakly-supervised hashing in kernel space," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3344–3351.

[19] L. Zhu, J. Shen, and L. Xie, "Topic hypergraph hashing for mobile image retrieval," in *Proc. ACM Int. Conf. Multimedia (MM)*, Brisbane, QLD, Australia, 2015, pp. 843–846.

[20] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. Joint Conf. Artif. Intell. (IJCAI)*, Barcelona, Spain, 2011, pp. 1360–1365.

[21] D. Zhang, F. Wang, and L. Si, "Composite hashing with multiple information sources," in *Proc. ACM Int. Conf. Inf. Retrieval (SIGIR)*, Beijing, China, 2011, pp. 225–234.

[22] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM Int. Conf. Manag. Data (SIGMOD)*, New York, NY, USA, 2013, pp. 785–796.

[23] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.

[24] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia (MM)*, Florence, Italy, 2010, pp. 251–260.

[25] L. Xie, P. Pan, and Y. Lu, "A semantic model for cross-modal and multi-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval (ICMR)*, Dallas, TX, USA, 2013, pp. 175–182.

[26] Y. Yang *et al.*, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[27] Y. Yan, G. Liu, S. Wang, J. Zhang, and K. Zheng, "Graph-based clustering and ranking for diversified image search," *Multimedia Syst.*, pp. 1–12, Sep. 2014, doi: 10.1007/s00530-014-0419-4.

[28] L. Zhang *et al.*, "Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1408–1419, Aug. 2014.

[29] V. Chandrasekhar *et al.*, "Transform coding of image feature descriptors," in *Proc. SPIE Conf. Vis. Commun. Image Process. (VCIP)*, San Jose, CA, USA, 2009, pp. 1–10.

[30] S. S. Tsai *et al.*, "Location coding for mobile image retrieval," in *Proc. Int. Conf. Mobile Multimedia Commun. (ICMMC)*, London, U.K., 2009, p. 8.

[31] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 37–45.

[32] L. Gao, J. Song, F. Zou, D. Zhang, and J. Shao, "Scalable multimedia retrieval by deep learning hashing with relative similarity learning," in *Proc. ACM Int. Conf. Multimedia (MM)*, Brisbane, QLD, Australia, 2015, pp. 903–906.

[33] J. Wang, S. Kumar, and S.-F. Chang, "Sequential projection learning for hashing with compact codes," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, 2010, pp. 1127–1134.

[34] R. Salakhutdinov and G. Hinton, "Semantic hashing," *Int. J. Approx. Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[35] Y. Liu, F. Wu, Y. Yang, Y. Zhuang, and A. G. Hauptmann, "Spline regression hashing for fast image search," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4480–4491, Oct. 2012.

[36] D. Zhang, J. Wang, D. Cai, and J. Lu, "Self-taught hashing for fast similarity search," in *Proc. ACM Int. Conf. Inf. Retrieval (SIGIR)*, Geneva, Switzerland, 2010, pp. 18–25.

[37] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, USA, 2011, pp. 1–8.

[38] J. Song, L. Gao, Y. Yan, D. Zhang, and N. Sebe, "Supervised hashing with pseudo labels for scalable multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia (MM)*, Brisbane, QLD, Australia, 2015, pp. 827–830.

[39] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 2083–2090.

[40] S. Kim, Y. Kang, and S. Choi, "Sequential spectral learning to hash with multiple representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, 2012, pp. 538–551.

[41] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[42] F. Zou *et al.*, "Compact image fingerprint via multiple kernel hashing," *IEEE Trans. Multimedia*, vol. 17, no. 7, pp. 1006–1018, Jul. 2015.

[43] L. Liu, M. Yu, and L. Shao, "Multiview alignment hashing for efficient image search," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 956–966, Mar. 2015.

[44] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. AAAI Conf. Artif. Intell.*, Phoenix, AZ, USA, 2016, pp. 1969–1976.

[45] L. Gao *et al.*, "Optimal graph learning with partial tags and multiple features for image and video annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 4371–4379.

[46] Y. Yang *et al.*, "Local image tagging via graph regularized joint group sparsity," *Pattern Recognit.*, vol. 46, no. 5, pp. 1358–1368, 2013.

[47] Y. Yang *et al.*, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[48] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas, "Image retrieval via probabilistic hypergraph ranking," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Francisco, CA, USA, 2010, pp. 3376–3383.

[49] Y. Gao *et al.*, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[50] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 4290–4303, Sep. 2012.

[51] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, vol. 1. Berkeley, CA, USA, 1967, pp. 281–297.

[52] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[53] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2, 1-norm regularized discriminative feature selection for unsupervised learning," in *Proc. Joint Conf. Artif. Intell. (IJCAI)*, Barcelona, Spain, 2011, pp. 1589–1594.

[54] F. Nie, H. Huang, X. Cai, and C. H. Q. Ding, "Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2010, pp. 1813–1821.

[55] K. Takeuchi, K. Ishiguro, A. Kimura, and H. Sawada, "Non-negative multiple matrix factorization," in *Proc. Joint Conf. Artif. Intell. (IJCAI)*, Beijing, China, 2013, pp. 1713–1720.

[56] X. Cai, F. Nie, W. Cai, and H. Huang, "Heterogeneous image features integration via multi-modal semi-supervised learning model," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, 2013, pp. 1737–1744.

[57] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. Joint Conf. Artif. Intell. (IJCAI)*, Beijing, China, 2013, pp. 2598–2604.

[58] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[59] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2003, pp. 321–328.

[60] T.-S. Chua *et al.*, "NUS-WIDE: A real-world Web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval (CIVR)*, Santorini, Greece, 2009, p. 48.

[61] Y. Zhen and D.-Y. Yeung, "A probabilistic model for multimodal hash function learning," in *Proc. ACM Int. Conf. Knowl. Disc. Data Min. (KDD)*, Beijing, China, 2012, pp. 940–948.

[62] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 2083–2090.

[63] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu, "Multimodal graph-based reranking for Web image search," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4649–4661, Nov. 2012.

[64] J. Xu, V. Singh, Z. Guan, and B. S. Manjunath, "Unified hypergraph for image ranking in a multimodal context," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 2333–2336.

[65] L. Zhu, J. Shen, H. Jin, R. Zheng, and L. Xie, "Content-based visual landmark search via multimodal hypergraph learning," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2756–2769, Dec. 2015.

[66] M. Avriel, *Nonlinear Programming: Analysis and Methods*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1976.