Supervised Matrix Factorization for Cross-Modality Hashing

Hong Liu[†‡]**, Rongrong Ji**[†‡∗]**, Yongjian Wu**[♮]**, Gang Hua**[♭]

[†]Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, 361005, China
[‡]School of Information Science and Engineering, Xiamen University, 361005, China
[♮]BestImage, Tencent Technology (Shanghai) Co.,Ltd, China
[♭]Microsoft Research Asia, Beijing, China
lynnliuxmu@outlook.com, rrji@xmu.edu.cn, littlekenwu@tencent.com,
ganghua@gmail.com

Abstract

Matrix factorization has been recently utilized for the task of multi-modal hashing for cross-modality visual search, where basis functions are learned to map data from different modalities to the same Hamming embedding. In this paper, we propose a novel cross-modality hashing algorithm termed Supervised Matrix Factorization Hashing (SMFH) which tackles the multi-modal hashing problem with a collective non-negative matrix factorization across the different modalities. In particular, SMFH employs a well-designed binary code learning algorithm to preserve the similarities among multi-modal original features through a graph regularization. At the same time, semantic labels, when available, are incorporated into the learning procedure. We conjecture that all these would facilitate to preserve the most relevant information during the binary quantization process, and hence improve the retrieval accuracy. We demonstrate the superior performance of SMFH on three cross-modality visual search benchmarks, *i.e.*, the PASCAL-Sentence, Wiki, and NUS-WIDE, with quantitative comparison to various state-of-the-art methods [Kumar and Udupa, 2011; Rastegari *et al.*, 2013; Zhang and Li, 2014; Ding *et al.*, 2014].

1 Introduction

Cross-modality retrieval has been a fundamental problem in several emerging applications including visual search, machine translation, and text mining [Bronstein *et al.*, 2010; Masci *et al.*, 2014; Rasiwasia *et al.*, 2010; Costa Pereira *et al.*, 2014]. In a typical scenario of cross-modality retrieval, a query comes from one modality, *e.g.*, text, while the returned results come from another modality, *e.g.*, image. To achieve this goal, a typical solution adopted by most existing works is to embed data samples of different modalities into a common low-dimensional space. By doing so, both the query and returns can be well aligned to capture their cross-modality similarities for retrieval [Costa Pereira *et al.*, 2014; Wang *et al.*, 2014].

Recently, both unsupervised and supervised hashing techniques have been investigated for cross-modality retrieval due to their prominent efficiency. For instance, Bronstein *et al.* proposed a Cross-Modality Similarity Search Hashing (CMSSH) algorithm by using eigen-decomposition and boosting. Both Cross-View Hashing (CVH) [Kumar and Udupa, 2011] and Inter-Media Hashing (IMH) [Song *et al.*, 2013] extended the classic Spectral Hashing approach [Weiss *et al.*, 2009] to the scenario of cross-modality retrieval. For another instance, Co-Regularized Hashing (CRH) [Zhen and Yeung, 2012] and Heterogeneous Translated Hashing (HTH) [Wei *et al.*, 2014] further deal with the cross-modality hashing under a co-regularized boosting framework.

In [Rastegari *et al.*, 2013], Predictable Dual-View Hashing (PDH) was proposed to learn the discriminative hash functions via a max-margin formulation with an iterative optimization algorithm. In [Ding *et al.*, 2014], Collective Matrix Factorization Hashing (CMFH) was proposed to formulate the joint learning of cross-modality binary codes as a collective matrix factorization problem. In [Zhang and Li, 2014], Supervised Multi-modal Hashing (SMH) was proposed to integrate semantic labels to improve the performance of hash function learning in the respective modalities.

While promising progress has been made, it remains as an open problem to capture the multi-modal similarities among data samples, as well as to preserve such similarities in a produced binary code (Hamming) space. As mentioned before, collective factorization [Ding *et al.*, 2014] and supervised hashing [Zhang and Li, 2014] have demonstrated outperformance on respective problems, *a.k.a.*, for cross-modality retrieval and for single-modality binary code learning [Liu *et al.*, 2015; Mukherjee *et al.*, 2015; Lin *et al.*, 2015]. The intuition of our work is to combine the merits of supervised hashing [Liu *et al.*, 2012; Zhang and Li, 2014] over the state-of-the-art cross-modality retrieval schemes [Ding *et al.*, 2014].

However, the integration of both approaches towards supervised cross-modality hashing is not an easy task. In one aspect, it is hard to optimize the discrete Hamming distances. In the other aspect, the complexity of the existing matrix factorization hashing and supervised cross-modality hashing are
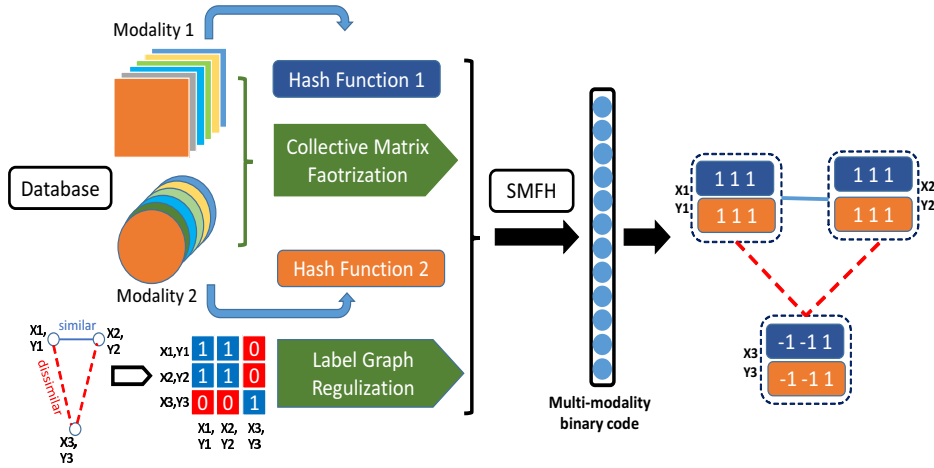
___

[∗]Corresponding author.

Figure 1: The Framework of our proposed Supervised Matrix Factorization Hashing (SMFH).

very high, which are the square of the training set size [Zhang and Li, 2014; Liu *et al.*, 2015; Mukherjee *et al.*, 2015] and cannot be easily scaled up to massive training data.

In this paper, we propose a novel cross-modality hashing method, dubbed the name Supervised Matrix Factorization Hashing (SMFH), which addresses the above challenges under a graph-regularized, collective non-negative matrix factorization framework. The first contribution we make is a hybrid regularization method to model the hash function learning, which integrates graph regularization into the collective non-negative matrix factorization. The second contribution we make is a supervised collective non-negative matrix factorization scheme, which leverages semantic labels to refine the graph regularizer during the step of graph construction. It ensures the learned binary codes to preserve the semantic similarities among data within multiple modalities.

Besides the two contributions mentioned above, we also propose an optimization algorithm to solve the objective function designed for SMFH, which works under an iterative updating procedure with stochastic sampling. This strategy can reach a training time reduction of supervised cross-modality hashing. Fig. 1 shows the overall framework of the proposed SMFH scheme. We conduct extensive experiments in cross-modality visual search, *i.e.*, using text queries to retrieve relevant images and vice versa, on three widely used benchmarks including, PASCAL-Sentence, Wiki and NUS-WIDE. We demonstrate the superior performance of SMFH over a group of state-of-the-art cross-modality hashing methods including CVH [Kumar and Udupa, 2011], PDH [Rastegari *et al.*, 2013], SMH [Zhang and Li, 2014], and CMFH [Ding *et al.*, 2014].

The rest of this paper is organized as follows: in Section 2, we present our SMFH approach in depth. Section 3 shows extensive experiments conducted on three benchmark datasets. Finally, we draw our conclusions in Section 4 and discuss the future work.

2 Supervised Matrix Factorization Hashing

In this section, we describe the proposed supervised cross-modality hashing algorithm. Without loss of generality, we take bi-modal hashing for instance, which can be easily extended to the scenario of multi-modality hashing.

2.1 Collective Factorization for Cross-Modality Hashing

Non-negative Matrix Factorization (NMF) is a matrix decomposition algorithm that focuses on learning low-rank representation. We define a non-negative data matrix $\mathbf{X} = [x_1, ..., x_N] \in \mathbb{R}^{d \times N}$, where N is the number of samples, d is the feature dimension, and x_i is the i-th sample. Non-negative matrix factorization aims at finding two non-negative factors $\mathbf{U} = [u_1, ..., u_r] \in \mathbb{R}^{d \times r}$ and $\mathbf{Y} = [v_1, ..., v_N] \in \mathbb{R}^{r \times N}$, $r << d$, whose product can approximate $\mathbf{X}$, *i.e.*,

$$\mathbf{X} \approx \mathbf{UY}. \qquad (1)$$

The squared Frobenius norm of the difference between two matrices is commonly used as the cost function to measure the approximation quality [Cai *et al.*, 2011; Liu *et al.*, 2013], which is defined as:

$$\min_{\mathbf{U},\mathbf{Y}} \| \mathbf{X} - \mathbf{UY} \|_F^2, \ s.t. \ \mathbf{U} \geqslant 0, \mathbf{Y} \geqslant 0, \qquad (2)$$

where $\| \cdot \|_F$ is the Frobenius norm of the matrix.

However, when facing cross-modality data, it is expected that the matrix $\mathbf{Y}$ is a new multi-modality representation of the data matrices $\mathbf{X}^1$ and $\mathbf{X}^2$ in a common low-dimensional space. It aims to learn two base $\mathbf{U}^1 \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{U}^2 \in \mathbb{R}^{d_2 \times r}$ to produce hash codes, which map bi-modal feature matrices into a r-dimensional binary code matrix $\mathbf{Y} \in \{0, 1\}^{r \times N}$, where r is the number of hash bits. Ideally, $\mathbf{Y}$ should reveal the hidden semantics shared by different modalities.

Correspondingly, the learning of the hash code is done via:

$$\min_{\mathbf{Y},\mathbf{U}^1,\mathbf{U}^2} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2, \qquad (3)$$

where λ_t is the weight coefficient of this modality satisfying $\sum_{t=1}^{2} \lambda_t = 1$. Intuitively, for the 1-st modal data $\mathbf{X}^1$, the matrix

$\mathbf{U}^1$ and the matrix $\mathbf{Y}$ are learned by Eq. (3), which is the same to $\mathbf{X}^2$. $y_{ji} = 1$ means that the corresponding semantic representation $u^t_{\cdot j}$ is related to data $\mathbf{X}^t_{\cdot i}$ and vice versa.

2.2 Learning Supervised Hash Function

Towards learning supervised hash function, our goal is to preserve the semantic similarity among data points in the Hamming space. That says, the hash functions should enforce labeled data pair $l(e_i) = l(e_j)$ to have identical or similar binary codes, where $e_i = (x^1_i, x^2_i)$ is the i-th data sample on both modalities from the training set, and function $l(\cdot)$ returns the supervised labels of the data across different modalities.

Our task is to find a graph matrix that preserves the intrinsic geometric structure of the similarity from bi-modal data. To this end, we first construct an affinity graph to model the aforementioned semantic information. This is done by calculating the pairwise similarity via the inner product among the semantic labels.

More specifically, the similarity between the i-th data sample and the j-th data sample is defined as follow:

$$\mathbf{A}_{ij} = l(e_i)^T \cdot l(e_j). \tag{4}$$

Without loss of generality, we define a matrix $\mathbf{L_a} \in \{0, 1\}^{dl \times N}$, where each column of $\mathbf{L_a}$ is the label representation vector of each data sample, and each row represents each sample's category[1]. Subsequently, the similarity matrix can be presented as $\mathbf{A} = \mathbf{L_a}^T \cdot \mathbf{L_a}$.

Then, our goal can be formulated via a spectral graph learning problem from the label similarity matrix $\mathbf{A}$ as:

$$\min_{\mathbf{Y}} \frac{1}{2} \sum_{i,j=1}^{n} \|\mathbf{Y}_i - \mathbf{Y}_j\|^2 \mathbf{A}_{i,j} = Tr(\mathbf{Y}^T \mathbf{L} \mathbf{Y}), \tag{5}$$

where $\mathbf{L}$ is the Laplacian matrix for $\mathbf{A}$.

Given data in two modalities, a.k.a, $\mathbf{X}^1 = \{x^1_i \in \mathbb{R}^{d_1} | i = 1, ..., N\}$, $\mathbf{X}^2 = \{x^2_i \in \mathbb{R}^{d_2} | i = 1, ..., N\}$, and their pairwise semantic similarity $\mathbf{A}$, the goal of supervised hash function learning is to learn two basis matrices $\mathbf{U}^1$ and $\mathbf{U}^2$, together with the hash codes $\mathbf{Y}$ by the following objective function:

$$\min_{\mathbf{Y}, \mathbf{U}^1, \mathbf{U}^2} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2 + \alpha Tr(\mathbf{Y}^T \mathbf{L} \mathbf{Y}), \tag{6}$$

where α is a balance parameter, which can be seen as a regularizer for the above collective non-matrix factorization.

Intuitively, we learn coefficients of the optimal projection $\mathbf{W}^t$ by minimizing the error term between the linear hash function $\mathbf{H}^t(\mathbf{X}^t) = sgn(\mathbf{W}^{t^T} \mathbf{X}^t)$ and the hash codes $\mathbf{Y}$ as $\|\mathbf{Y} - \mathbf{H}^t(\mathbf{X}^t)\|^2$. Inspired by [He et al., 2015], label graph can be used as Laplacian Regularized in this error. Then, the objective function of hash function learning is written as follows:

$$\min_{\mathbf{H}^t(\mathbf{X}^t)} \|\mathbf{Y} - \mathbf{H}^t(\mathbf{X}^t)\|^2 + \eta Tr(\mathbf{H}^t(\mathbf{X}^t)\mathbf{L}\mathbf{H}^t(\mathbf{X}^t)^T), \tag{7}$$

where η is the balance parameter. This term is integrated with the term of supervised semantic similarity in the proposed graph-regularized collective matrix factorization. Then, the overall objective function is written as follows:

$$\min_{\mathbf{Y}, \mathbf{W}^t} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2 + \alpha Tr(\mathbf{Y} \mathbf{L} \mathbf{Y}^T)$$
$$+\mu \sum_{t=1}^{2} \{ \|\mathbf{Y} - \mathbf{H}^t(\mathbf{X}^t)\|_F^2 + \eta Tr(\mathbf{H}^t(\mathbf{X}^t)\mathbf{L}\mathbf{H}^t(\mathbf{X}^t)^T) \} \tag{8}$$
$$s.t. \ \mathbf{Y} \in \{0, 1\}^{r \times N}, \ \mathbf{Y}\mathbf{Y}^T = \mathbf{I}, \ \mathbf{U}^t \geq 0,$$

---

[1]In this paper, we assume that L_a is fully observed without missing labels. We can get the labels among points in many cases for missing labeled datas, i.e. classification technology, which makes our assumption reasonable.

where μ is a tradeoff parameter to control the weights between hash function approximation and the proposed graph-regularized collective non-matrix factorization scheme.

2.3 Optimization

Directly minimizing the objective function in Eq. (8) is intractable due to the discrete constraint of $\mathbf{Y}$. To tackle this issue, we relax the discrete constraint from $\mathbf{Y} \in \{0, 1\}^{r \times n}$ to $0 \leq \mathbf{Y} \leq 1$. After that, it is still non-convex with respect to $\mathbf{Y}, \mathbf{U}^t$, and $\mathbf{W}^t$ jointly. This is further handled by using an alternating optimization, i.e., updating one variable while fixing the rest two until convergence. Due to the scale of the label similarity matrix, it is inefficient to store and compute it during optimization. To solve this problem, we propose a random sampling method, which uses a sampled sub-graph to replace the complete similarity matrix. The detailed optimization procedure is presented as follows:

(1) **Fix $\mathbf{W}^t$ and $\mathbf{Y}$, then update $\mathbf{U}^t$.** The corresponding sub-problem is

$$\min_{\mathbf{U}^1, \mathbf{U}^2} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2 \tag{9}$$
$$s.t. \ \mathbf{U}^1 \geq 0, \ \mathbf{U}^2 \geq 0, \ \lambda_1 + \lambda_2 = 1.$$

Here we learn two basis matrices $\mathbf{U}^1$ and $\mathbf{U}^2$, which turns this sub-problem to be a traditional NMF problem for each modality.

By directly optimizing Eq. (9) within each modality respectively, we solve the new objective function through Lagrange multiplier $\Psi = [\psi_{ik}]$ with constant $u^t_{ij} \geq 0$:

$$O_1(\mathbf{U}^t) = Tr(\mathbf{X}^t \mathbf{X}^{t^T}) - 2Tr(\mathbf{X}^t \mathbf{Y}^T \mathbf{U}^{t^T})$$
$$+ Tr(\mathbf{U}^t \mathbf{Y} \mathbf{Y}^T \mathbf{U}^{t^T}) + Tr(\Psi \mathbf{U}^{t^T}). \tag{10}$$

We derive the partial derivatives with respect to $\mathbf{U}^t$ as :

$$\frac{\partial O_1}{\partial \mathbf{U}^t} = -2\mathbf{X}^t \mathbf{Y}^T + 2\mathbf{U}^t \mathbf{Y} \mathbf{Y}^T + \Psi. \tag{11}$$

Then by Karush-Kuhn-Tucker (KKT) conditions, we update $\mathbf{U}^t$ via the following procedure:

$$u^t_{ij} \leftarrow u^t_{ij} \frac{(\mathbf{X}^t \mathbf{Y}^T)_{ij}}{(\mathbf{U}^t \mathbf{Y} \mathbf{Y}^T)_{ij}}. \tag{12}$$

Intuitively, $\mathbf{U}^t$ is non-negative after updating.

(2) **Fix $\mathbf{U}^t$ and $\mathbf{W}^t$, then update $\mathbf{Y}$.** We then fix $\mathbf{U}^t$ and $\mathbf{W}^t$, the updating of $\mathbf{Y}$ subsequently refers to:

$$\min_{\mathbf{Y}} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2 + \alpha Tr(\mathbf{Y}^T \mathbf{L} \mathbf{Y})$$
$$+\mu \sum_{t=1}^{2} \|\mathbf{Y} - \mathbf{W}^{t^T} \mathbf{X}^t\|_F^2 + \beta \|\mathbf{Y}\mathbf{Y}^T - \mathbf{I}\|_F^2 \tag{13}$$
$$s.t. \ 0 \leq \mathbf{Y} \leq 1.$$

However, the scale of the semantic matrix is extremely large, which needs huge storage cost and makes the Eq. (12) hard to optimize during each iteration. To tackle this problem, we randomly sample parts of the original label similarity matrix A, which can approximate the graph regularization. We define a sampling matrix $\mathbf{S} \in \{0, 1\}^{N \times m}$ during each iteration, where m is the number of sampling points with $m << N$. Then the hash code of sampled data can be represented by $\hat{\mathbf{Y}} = \mathbf{Y}^T \mathbf{S} \in \{0, 1\}^{d_t \times m}$, and the sampled label representation can be presented by $\hat{\mathbf{L}_a} = \mathbf{L_a}^T \mathbf{S} \in \{0, 1\}^{dl \times m}$. By using the above sampling, this sub-problem of Eq. (12) can be rewritten as:

$$\min_{\mathbf{Y}} \sum_{t=1}^{2} \lambda_t \|\mathbf{X}^t - \mathbf{U}^t \mathbf{Y}\|_F^2 + \alpha Tr(\hat{\mathbf{Y}}^T \hat{\mathbf{L}} \hat{\mathbf{Y}})$$
$$+\mu \sum_{t=1}^{2} \|\mathbf{Y} - \mathbf{W}^{t^T} \mathbf{X}^t\|_F^2 + \beta \|\mathbf{Y}\mathbf{Y}^T - \mathbf{I}\|_F^2 \tag{14}$$
$$s.t. \ 0 \leq \mathbf{Y} \leq 1.$$

Algorithm 1 Supervised Matrix Factorization Hashing

Input: Training data points in two modalities $\mathbf{X}^1$ and $\mathbf{X}^2$, the corresponding pairwise semantic similarity matrix $\mathbf{A}$, the number of support samples m, and the number of hash bits r.

Output: The hash codes $\mathbf{Y}$ for training data and the projection coefficient matrix $\mathbf{W}^t$.

1: Initialize $\mathbf{W}^t$, $\mathbf{U}^t$ and $\mathbf{Y}$ by random matrices, $t = 1, 2$.
2: **repeat**
3:     Fixing $\mathbf{W}^t$ and $\mathbf{Y}$, update $\mathbf{U}^t$ by Eq. (12);
4:     Uniformly and randomly select m sample pairs from training data.
5:     Fixing $\mathbf{U}^t$ and $\mathbf{W}^t$, update $\mathbf{Y}$ by Eq. (17);
6:     Fixing $\mathbf{U}^t$ and $\mathbf{Y}$, update $\mathbf{W}^t$ by Eq. (18);
7: **until** convergence

---

where $\hat{\mathbf{L}}$ is the Laplacian matrix for $\hat{\mathbf{A}} = \hat{\mathbf{L_a}}^T \cdot \hat{\mathbf{L_a}}$.

Since solving such a constraint is not convenient, we further relax it to $\mathbf{Y} \geq 0$ and normalize $\mathbf{Y}$ after factorization in each iteration. Let Φ be the Lagrange Multiplier for the new constraint. The Lagrange term could be written as follows:

$$
\begin{aligned}
O_2(\mathbf{Y}) = &\sum_{t=1}^{2} \lambda_t \{ Tr(\mathbf{X}^t \mathbf{X}^{t^T}) - 2Tr(\mathbf{X}^t \mathbf{Y}^T \mathbf{U}^{t^T}) \\
&+ Tr(\mathbf{U}^t \mathbf{Y} \mathbf{Y}^T \mathbf{U}^{t^T}) \} + \alpha Tr(\hat{\mathbf{Y}} \hat{\mathbf{L}} \hat{\mathbf{Y}}^T) \\
&+ \mu \sum_{t=1}^{2} \{ Tr(\mathbf{Y} \mathbf{Y}^T) - 2Tr(\mathbf{Y} \mathbf{X}^{t^T} \mathbf{W}^t) \\
&+ Tr(\mathbf{W}^{t^T} \mathbf{X}^t \mathbf{X}^{t^T} \mathbf{W}^t) \} + \beta \{ Tr(\mathbf{Y} \mathbf{Y}^T \mathbf{Y} \mathbf{Y}^T) \\
&- 2Tr(\mathbf{Y} \mathbf{Y}^T) + const \} + Tr(\Phi \mathbf{Y}^T).
\end{aligned}
\tag{15}
$$

Using the KKT conditions, we have:

$$
\begin{aligned}
\frac{\partial O_2}{\partial \mathbf{Y}} = &2 \sum_{t=1}^{2} \lambda_t \{ -\mathbf{U}^{t^T} \mathbf{X}^t + \mathbf{U}^{t^T} \mathbf{U}^t \mathbf{Y} \} + 2\alpha \hat{\mathbf{Y}} \hat{\mathbf{L}} \mathbf{S}^T \\
&+ 2\mu \sum_{t=1}^{2} \{ \mathbf{Y} - \mathbf{W}^{t^T} \mathbf{X}^t \} + 4\beta \{ \mathbf{Y} \mathbf{Y}^T \mathbf{Y} - \mathbf{Y} \} + \Phi = 0. \\
&s.t. \ \psi_{ij} \mathbf{Y}_{ij} = 0, \mathbf{Y} \geq 0.
\end{aligned}
\tag{16}
$$

Eq. (14) can be solved by the following updating rule:

$$
y_{ij} \leftarrow y_{ij} \cdot
$$
$$
\frac{(\sum_{t=1}^{2} \lambda_t \mathbf{U}^{t^T} \mathbf{X}^t + \alpha \hat{\mathbf{Y}} \hat{\mathbf{A}} \mathbf{S}^T + \mu \sum_{t=1}^{2} \mathbf{W}^{t^T} \mathbf{X}^t + 4\beta \mathbf{Y})_{ij}}{(\sum_{t=1}^{2} \lambda_t \mathbf{U}^{t^T} \mathbf{U}^t \mathbf{Y} + \alpha \hat{\mathbf{Y}} \hat{\mathbf{D}} \mathbf{S}^T + \mu \sum_{t=1}^{2} \mathbf{Y} + 4\beta \mathbf{Y}(\mathbf{Y}^T \mathbf{Y}))_{ij}}.
\tag{17}
$$

$\hat{\mathbf{D}}$ is the diagonal matrix with entries of column sums of $\hat{\mathbf{A}}$.

(3) **Fix $\mathbf{U}^t$ and $\mathbf{Y}$, then update $\mathbf{W}^t$.** This last sub-problem finds the best projection coefficient $\mathbf{W}^t$ by minimizing Eq. (7) for the t-th modality as the Laplacian Regularized Least squares algorithm [He *et al.*, 2015], resulting in a closed-form solution:

$$
\mathbf{W}^t = (\mathbf{X}^t \mathbf{X}^{t^T} + \eta(\mathbf{X}^t \mathbf{S}) \hat{\mathbf{L}} (\mathbf{S}^T \mathbf{X}^{t^T}) + \gamma \mathbf{I})^{-1} \mathbf{X}^t \mathbf{Y}^T.
\tag{18}
$$

We summarize the whole procedure of the proposed SMFH in Algorithm 1.

Times Complexity: The main time consumption of the proposed SMFH is the matrix factorization, its complexity is $O\big((ndr + (n + d)(r^2 + r) + m^2 r)t \big)$, where t is the number of iterations. Since $r, m << n$, the overall complexity is $O\big(n(d+r)rt\big)$, which is linear to the size of the training data. In practice, the proposed SMFH is much faster than most cross-modality hashing, *i.e.* PDH and CMFH, whose training time is rounded to the competing cross-moality hashing method SMH. Although the hash codes of training data is obtained by minimizing Eq. (8), it cannot be directly applied to the case of the out-of-sample query. For such out-of-sample query, we use the hash function learned by Eq. (8) to generate the corresponding binary code. Then in online search, the time complexity for each modality is constant as $O(dr)$.

2.4 Extension to Multi-Modality Search

It is quite intuitive to extend SMFH in Eq. (8) from bi-modal to multiple modalities, that is:

$$
\min_{\mathbf{Y}, \mathbf{W}^t} \sum_{t=1}^{n_t} \lambda_t \| \mathbf{X}^t - \mathbf{U}^t \mathbf{Y} \|_F^2 + \alpha Tr(\mathbf{Y} \mathbf{L} \mathbf{Y}^T)
\tag{19}
$$
$$
+ \mu \sum_{t=1}^{n_t} \{ \| \mathbf{Y} - \mathbf{H}^t(\mathbf{X}^t) \|_F^2 + \eta Tr(\mathbf{H}^t(\mathbf{X}^t) \mathbf{L} \mathbf{H}^t(\mathbf{X}^t)^T) \}
$$
$$
s.t. \ \mathbf{Y} \in \{0, 1\}^{r \times n}, \ \mathbf{U}^t \geq 0,
$$

where $\sum_{t=1}^{n_t} \lambda_t = 1$. It is convenient to adopt Algorithm 1 to minimize the objective function in Eq. (8). An alternating optimization strategy can also be used here. In particular, the variable $\mathbf{U}^t$ and $\mathbf{W}^t$ can be directly got through Eq. (12) and Eq. (17), respectively. And finally, the variable $\mathbf{Y}$ can be learned by the new formulation as follows:

$$
y_{ij} \leftarrow y_{ij} \cdot
$$
$$
\frac{(\sum_{t=1}^{n_t} \lambda_t \mathbf{U}^{t^T} \mathbf{X}^t + \alpha \hat{\mathbf{Y}} \hat{\mathbf{A}} \mathbf{S}^T + \mu \sum_{t=1}^{n_t} \mathbf{W}^{t^T} \mathbf{X}^t + 4\beta \mathbf{Y})_{ij}}{(\sum_{t=1}^{n_t} \lambda_t \mathbf{U}^{t^T} \mathbf{U}^t \mathbf{Y} + \alpha \hat{\mathbf{Y}} \hat{\mathbf{D}} \mathbf{S}^T + \mu \sum_{t=1}^{n_t} \mathbf{Y} + 4\beta \mathbf{Y}(\mathbf{Y}^T \mathbf{Y}))_{ij}}.
\tag{20}
$$

3 Experiments

Quantitative experiments are conducted to validate the advantages of the proposed cross-modality hashing algorithm on three widely-used benchmark, *i.e.*, **PASCAL-Sentence**[2], **Wiki**[3] and **NUS-WIDE**[4].

The **PASCAL-Sentence** dataset contains 1,000 images that are divided into 20 categories. Each image is represented by a 269-dimensional visual feature extracted by a collections detectors. A 2,790-dimensional textual feature is extracted using the bag-of-words representation with WordNet [Farhadi *et al.*, 2010]. For this dataset, 800 image-sentence pairs are randomly sampled as the training set and the remaining for query testing.

The **Wiki** dataset contains 2,866 documents, where the image-text pairs are fully annotated with 10 semantic categories. Each image is represented as a 128-dimensional bag-of-visual-words feature. Each document is represented as a 10-dimensional topical feature using Latent Dirichlet Allocation [Blei *et al.*, 2003]. For the Wiki dataset, we randomly select 75% image-text pairs for training and the rest for query testing.

The **NUS-WIDE** dataset contains 269,648 images with 81 concepts crawled from Flickr. We select 186,577 labeled image-text pairs according to the top 10 largest concepts as adopted in [Hu *et*

---

[2]http://vision.cs.uiuc.edu/pascal-sentences/

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

[4]http : //lms.comp.nus.edu.sg/research/NUSWIDE.htm

Table 1: The *m*AP and Precision Comparison Using Hamming Ranking on Two Benchmark with Different Hash Bits.

Task	Methods	Wiki					NUS-WIDE				
		*m*AP			Precision@100		*m*AP			Precision@100	
		32	64	128	32	64	32	64	128	32	64
Task 1	CVH	0.2053	0.1872	0.2039	0.1504	0.1324	0.4480	0.4184	0.4012	0.4606	0.4281
	PDH	0.2034	0.2047	0.2133	0.1765	0.1725	0.5008	0.5078	0.5336	0.4885	0.5018
	CMFH	0.5947	0.6063	0.6131	0.5492	0.5649	0.3807	0.3787	0.3663	0.3764	0.3750
	SMH	0.3831	0.4032	0.4171	0.3135	0.3400	**0.5978**	0.6162	0.6195	**0.6007**	0.6194
	SMFH	**0.6039**	**0.6602**	**0.6658**	**0.5581**	**0.6246**	0.5462	**0.6633**	**0.6247**	0.5478	**0.6757**
Task 2	CVH	0.1660	0.1479	0.1572	0.1280	0.1170	0.4592	0.4260	0.4021	0.4700	0.4333
	PDH	0.2442	0.2360	**0.2685**	0.2058	0.1945	0.5129	0.5260	0.5377	0.4988	0.5224
	CMFH	0.2081	0.2111	0.2270	0.1691	0.1678	0.3818	0.3774	0.3664	0.3787	0.3766
	SMH	0.2301	0.2503	0.2570	0.1953	0.2186	0.5823	0.6020	0.6089	0.5829	0.6039
	SMFH	**0.2516**	**0.2581**	0.2496	**0.2168**	**0.2330**	**0.5938**	**0.6325**	**0.6175**	**0.5985**	**0.6380**

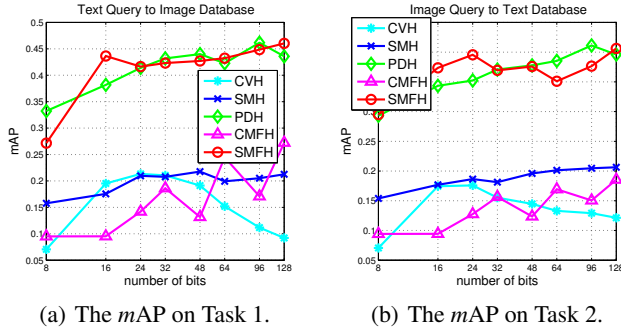(a) The *m*AP on Task 1.   (b) The *m*AP on Task 2.

Figure 2: The *m*AP curves on PASCAL-Sentence.

al., 2014; Zhang and Li, 2014]. In this dataset, images are represented by a 500-dimensional bag-of-visual-words feature, and its corresponding tags are represented by a 1,000-dimensional bag-of-words feature. We choose 6,577 image-text pairs from this database as query and the remaining to form the dataset for training.

Compared Methods: We evaluate the cross-modality retrieval task via: (1) the text-to-image side, termed **Task 1**, and (2) the image-to-text side, termed **Task 2**. In both sides, the proposed **SMFH** is compared against four state-of-the-art methods: Cross-View Hashing (**CVH**) [Kumar and Udupa, 2011], Supervised MultiModal Hashing (**SMH**) [Zhang and Li, 2014], Predictable Dual-view Hashing (**PDH**) [Rastegari *et al.*, 2013] and Collective Matrix Factorization Hashing (**CMFH**) [Ding *et al.*, 2014]. Except CVH, the source codes of the rest methods are available publicly, and all their parameters' setting are used as what their papers presented. All our experiments were run on a workstation with a 3.60GHz Intel Core I5-4790 CPU and 16GB RAM.

Evaluation Protocols: The quantitative performance is evaluated by using the mean Average Precision (*m*AP) with top 100 ranking list. We also consider other three evaluation protocols, *i.e.*, precision at top-100 positions (Precision@100), Recall curves at top-K and Precision curves at top-K.

Parameter Settings: SMFH has five essential parameters in Eq. (13), *i.e.*, λ_1, λ_2, α, μ, η and m. The parameter λ_1 and λ_2 control the weights between two modalities, which are found to have little influence. In our experiments, we empirically set $\lambda_1 = 0.5$ for the image modality and $\lambda_2 = 0.5$ for the text modality. The parameter α holds the semantic similarity of the original space, which is set as a large number of 2. μ is a trade-off parameter, which is set as 25 on the two datasets. During each iteration, the number of sampling

points is set as 800 for PASCAL-Sentence, $1,000$ for Wiki, and $2,000$ for the NUS-WIDE. At the last part of this section, we will analyze the relation of the parameter m, and show the convergence result. The three regularization parameters γ, β, and η are set to a small number 0.001 in all the experiments.

Quantitative Results: Fig.2 shows the *m*AP results on PASCAL-Sentence dataset with different bits on both retrieval tasks. SMFH has achieved remarkable *m*AP scores, especially when hash bit is larger than 32. Comparing to the second best scheme, SMFH has achieved 7.1% *m*AP improvement for the Task 1 and 6.5% improvement for the Task 2.

Then, we evaluate the proposed method on Wiki, as shown in the first row of Fig.3 and Tab.1, which demonstrate that SMFH has achieved superior performance on this benchmark for both text-to-image and image-to-text sides, both with a performance gain of more than 6%. The *m*AP results and Precision@100 results on Wiki are reported in Tab.1 under the setting of 32, 64, and 128 bits respectively. SMFH has achieved remarkable *m*AP and precision scores. Comparing with the state-of-the-art alogrithms, *i.e.*, [Kumar and Udupa, 2011; Rastegari *et al.*, 2013; Zhang and Li, 2014; Ding *et al.*, 2014], for the task of text-to-image retrieval, our SMFH has significant advantage on precision and *m*AP values with all bits, mainly due to the fact that the matrix factorization can successfully find better latent topic concepts from text. Meanwhile SMFH fully uses supervised label to improve the cross-modality retrieval. Fig.3 shows the comparison of precision curves and recall curves on Wiki when hash bit is 64.

Similar performance gains are observed on the large-scale NUS-WIDE, especially in the text-to-image retrieval. As shown in Tab.1 and the second row of Fig.3, SMFH achieves highest search accuracy. When the hash bit is 64, the precision of the top-100 Hamming ranking is over 60% by SMFH, which is much better than the state-of-the-arts [Ding *et al.*, 2014; Kumar and Udupa, 2011; Rastegari *et al.*, 2013; Zhang and Li, 2014]. Although, the *m*AP is not the highest when hash bit is 32 for text-to-image sides, SMFH maintain its the competitive advantage on higher hash bits. Nevertheless, the performance of Task 2 is at the second place, which also have competitive performance for cross-modality retrieval.

Tab.2 shows the results of training time comparing with different algorithms on different hash bits on three benchmarks, which contain the whole training set. PDH and CMFH are much higher than that of our proposed SMFH, with larger size of training data and higher feature dimensions. SMFH can get the better performance comparing with others, by fully use of the semantic information to enhance the performance with less training time.

We further study the influence of different sizes of the training set. As shown in Fig.4 (a) for Wiki, *m*AP results are shown when hash bit
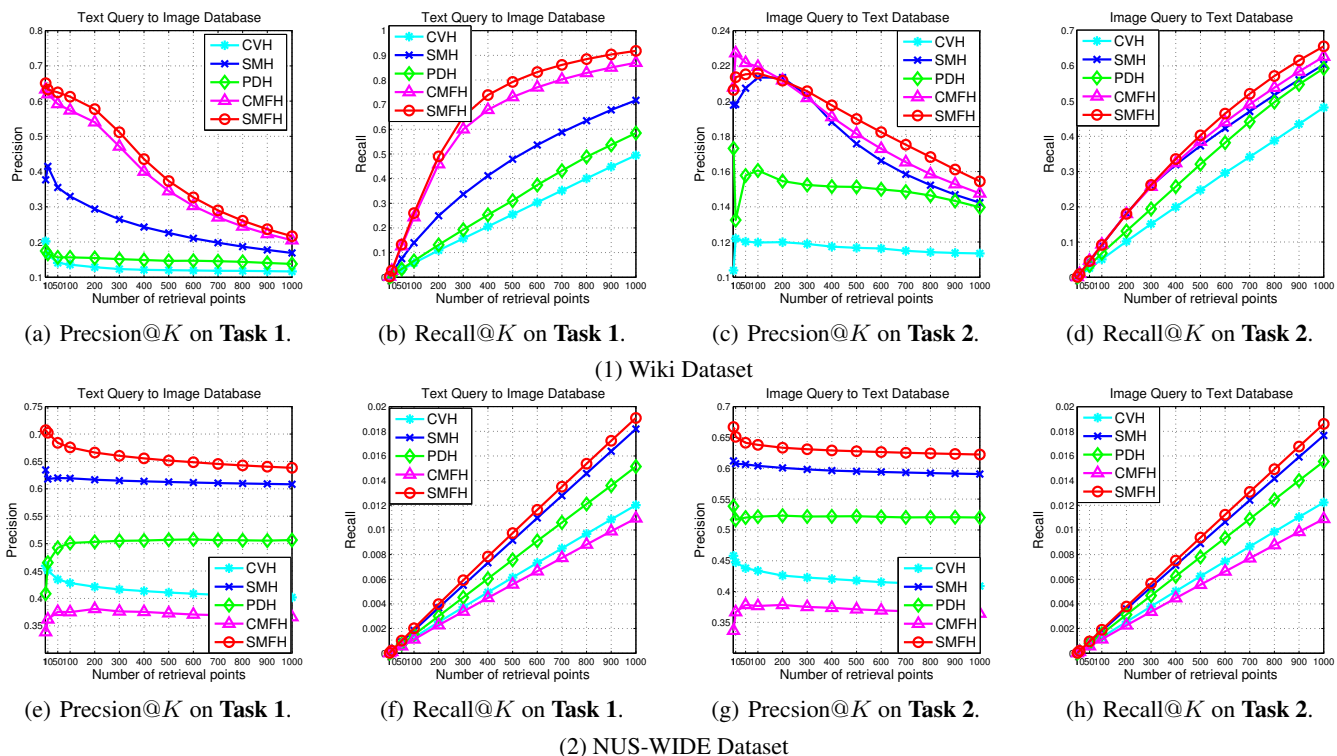
(a) Precsion@K on **Task 1**.   (b) Recall@K on **Task 1**.   (c) Precsion@K on **Task 2**.   (d) Recall@K on **Task 2**.

(1) Wiki Dataset

(e) Precsion@K on **Task 1**.   (f) Recall@K on **Task 1**.   (g) Precsion@K on **Task 2**.   (h) Recall@K on **Task 2**.

(2) NUS-WIDE Dataset

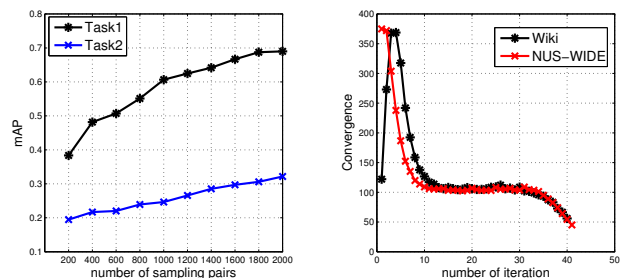Figure 3: The Precision curves and Recall curves of all the algorithms on both datasets when hash bit is 64.

Table 2: The Training Time (s) comparing with different algorithms on both datasets.

	PASCAL		Wiki		NUS-WIDE	
Methods	32	64	32	64	32	64
CVH	5.84	5.18	0.14	0.04	3.31	3.25
PDH	278.19	350.73	1.34	2.44	736.81	1461.71
CMFH	324.87	335.41	0.32	0.59	1579.1	1782.6
SMH	26.26	33.31	0.31	0.59	20.74	40.53
SMFH	5.89	6.16	0.56	0.71	73.94	88.04

(a) The mAP curves vs. number of sampling pairs.   (b) Convergence validation.

Figure 4: Parameter Analysis.

is 64, in which we vary the size of sampling from 200 to 2, 000. The performance of cross-modality retrieval consistently improve with the increasing of the sampling size for SMFH. Thus, we randomly choose about 100 image-tag pairs for each concept as the training set for convenience in optimization, which contain about 1, 000 pairs during each iteration. At last, we validate the convergence according to the sampling during each iteration. As shown in Fig.4 (b), when the size of sampling pairs is 1, 000, and the hash bit is 64, SMFH can quickly converge by using random sampling. The same conclusion holds on the NUS-WIDE dataset.

4 Conclusions

In this paper, we propose a novel hashing method termed Supervised Matrix Factorization Hashing (SMFH) for cross-modality visual search. We employ graph regularization to develop a collective matrix factorization based hashing framework, which can preserve the similarities among original features from different modalities into a produced Hamming space. Meanwhile, SMFH incorporates supervised label information to enhance the quantization quality of the learned binary codes. Furthermore, hashing and graph

regularization are integrated into a unified framework by means of joint hash function learning. In this framework, the given supervised labels can be leveraged to construct a label matrix, leading to more discriminative hash codes. Extensive experiments conducted on PASCAL-Sentence, Wiki, and NUS-WIDE benchmarks demonstrated the superior performance of SMFH over several state-of-the-art cross-modality hashing methods [Ding *et al.*, 2014; Kumar and Udupa, 2011; Rastegari *et al.*, 2013; Zhang and Li, 2014]. In the future, we would investigate large-scale discrete optimization techniques for the proposed SMFH.

5 Acknowledgement

References

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

[Bronstein *et al.*, 2010] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Proceedings of CVPR*, pages 3594–3601. IEEE, 2010.

[Cai *et al.*, 2011] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *TPAMI*, 33(8):1548–1560, 2011.

[Costa Pereira *et al.*, 2014] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535, 2014.

[Ding *et al.*, 2014] Guiguang Ding, Yuchen Guo, and Jile Zhou. Collective matrix factorization hashing for multimodal data. In *Proceedings of CVPR*, pages 2083–2090. IEEE, 2014.

[Farhadi *et al.*, 2010] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, pages 15–29. Springer, 2010.

[He *et al.*, 2015] Xiaofei He, Chiyuan Zhang, Lijun Zhang, and Xuelong Li. A-optimal projection for image representation. *TPAMI*, 38(5):1009–1015, 2015.

[Hu *et al.*, 2014] Yao Hu, Zhongming Jin, Hongyi Ren, Deng Cai, and Xiaofei He. Iterative multi-view hashing for cross media indexing. In *Proceedings of MM*, pages 527–536. ACM, 2014.

[Kumar and Udupa, 2011] Shaishav Kumar and Raghavendra Udupa. Learning hash functions for cross-view similarity search. In *Proceedings of IJCAI*, volume 22, page 1360, 2011.

[Lin *et al.*, 2015] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of CVPR*, pages 3864–3872, 2015.

[Liu *et al.*, 2012] Wei Liu, Jun Wang, Rongrong Ji, Yu-Gang Jiang, and Shih-Fu Chang. Supervised hashing with kernels. In *Proceedings of CVPR*, pages 2074–2081. IEEE, 2012.

[Liu *et al.*, 2013] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of SDM*, volume 13, pages 252–260. SIAM, 2013.

[Liu *et al.*, 2015] Li Liu, Mengyang Yu, and Ling Shao. Multiview alignment hashing for efficient image search. *TIP*, 24(3):956–966, 2015.

[Masci *et al.*, 2014] Jonathan Masci, Michael M Bronstein, Alexander M Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *TPAMI*, 36(4):824–830, 2014.

[Mukherjee *et al.*, 2015] Lopamudra Mukherjee, Sathya N Ravi, Vamsi K Ithapu, Tyler Holmes, and Vikas Singh. An nmf perspective on binary hashing. In *Proceedings of CVPR*, pages 4184–4192, 2015.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of MM*, pages 251–260. ACM, 2010.

[Rastegari *et al.*, 2013] Mohammad Rastegari, Jonghyun Choi, Shobeir Fakhraei, Daume Hal, and Larry Davis. Predictable dual-view hashing. In *Proceedings of ICML*, pages 1328–1336, 2013.

[Song *et al.*, 2013] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of SIGMOD*, pages 785–796. ACM, 2013.

[Wang *et al.*, 2014] Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. Effective multi-modal retrieval based on stacked auto-encoders. In *Proceedings of VLDB*, 2014.

[Wei *et al.*, 2014] Ying Wei, Yangqiu Song, Yi Zhen, Bo Liu, and Qiang Yang. Scalable heterogeneous translated hashing. In *Proceedings of SIGKDD*, pages 791–800. ACM, 2014.

[Weiss *et al.*, 2009] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *Proceedings of NIPS*, pages 1753–1760, 2009.

[Zhang and Li, 2014] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceedings of AAAI*, 2014.

[Zhen and Yeung, 2012] Yi Zhen and Dit-Yan Yeung. Co-regularized hashing for multimodal data. In *Proceeding of NIPS*, pages 1376–1384, 2012.