Cross-Modality Binary Code Learning via Fusion Similarity Hashing

Hong Liu[†], Rongrong Ji[†*], Yongjian Wu[‡], Feiyue Huang[‡], and Baochang Zhang[♮]
[†]Xiamen University, [‡]Tencent Technology (Shanghai) Co.,Ltd, [♮]Beihang University
lynnliu.xmu@gmail.com, rrji@xmu.edu.cn, littlekenwu@tencent.com
garyhuang@tencent.com, bczhang@buaa.edu.cn

Abstract

Binary code learning has recently been emerging topic in large-scale cross-modality retrieval. It aims to map features from multiple modalities into a common Hamming space, where the cross-modality similarity can be approximated efficiently via Hamming distance. To this end, most existing works learn binary codes directly from data instances in multiple modalities, which preserve both intra- and inter-modal similarities respectively. Few methods consider to preserve the "fusion similarity" among multi-modal instances instead, which can explicitly capture their heterogeneous correlation in cross-modality retrieval. In this paper, we propose a hashing scheme, termed Fusion Similarity Hashing (FSH), which explicitly embeds the graph-based fusion similarity across modalities into a common Hamming space. Inspired by the "fusion by diffusion", our core idea is to construct an undirected asymmetric graph to model the fusion similarity among different modalities, upon which a graph hashing scheme with alternating optimization is introduced to learn binary codes that embeds such fusion similarity. Quantitative evaluations on three widely used benchmarks, i.e., UCI Handwritten Digit, MIR-Flickr25K and NUS-WIDE, demonstrate that the proposed FSH approach can achieve superior performance over the state-of-the-art methods.

1. Introduction

Cross-modality visual search has been an emerging topic in computer vision communities recently [2, 20, 4, 24]. In a typical setting, instances in one modality, *e.g.*, text documents, are retrieved given a query from another modality, *e.g.*, image, and vice versa [29]. Ideally, the most similar instances to the query should occupy the top positions in the ranking list. Due to its low storage cost and fast retrieval speed, binary code learning, *a.k.a.*, hashing, has attracted much attention recently in cross-modality retrieval, which targets to find a low-dimensional Hamming space to efficiently preserve the cross-modality similarity.

Both unsupervised and supervised hashing schemes have been recently studied in cross-modality retrieval. For unsupervised hashing, both Cross-view hashing (CVH) [12] and Inter-Media Hashing (IMH) [23] are extended from Spectral Hashing [27] to fit the scenario of cross-modality retrieval. In [21], Predictable Dual-View Hashing (PDH) was proposed to learn two linear hash functions via a self-taught learning algorithm. Collective Matrix Factorization Hashing (CMFH) [5] aims to finding consistent hash codes from different views by collective matrix factorization. In [33], Latent Semantic Sparse Hashing (LSSH) was proposed to learn latent features from images and texts jointly with sparse coding, upon which hash codes are learned. Differently, supervised hashing embeds the label supervision into the Hamming space to improve the retrieval performance, for instance Co-Regularized Hashing (CRH) [31], Heterogeneous Translated Hashing (HTH) [26], Supervised Multi-Modal Hashing (SMH) [30], Quantized Correlation Hashing (QCH) [28], Semantics-Preserving Hashing (SePH)[14], and Supervised Matrix Factorization Hashing (SMFH) [15]. Although supervised hashing typically achieves superior performance, it is very labor intensive to obtain large-scale labels in many real-world applications.

In this paper, we focus on unsupervised hashing for cross-modality retrieval. To learn discriminative binary codes, it is essential to preserve the intra- and inter-modal similarities jointly in the common Hamming space produced. To this end, many existing cross-modality hashing, *i.e.*, Cross-view Hashing, Inter-Media Hashing, and Linear Cross-Modal Hashing [34] preserve both similarities in a *separated* manner, typically under a co-training setting that minimizes the intra- and inter- modal loss iteratively and respectively. However, most of them handle the problem of such two similarities preserving separately to learn the corresponding hash codes with a co-training algorithm. More importantly, such methods neglect to preserve the fusion similarity among multi-modalities data. We argue that such fusion similarity are more important for measuring the

*Corresponding author.

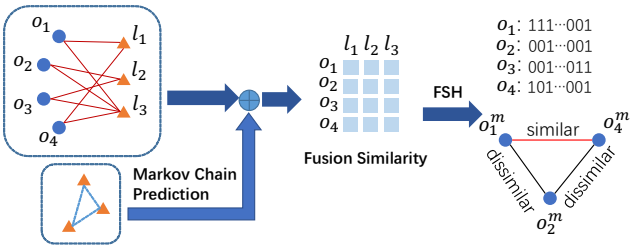


Figure 1. The Framework of our proposed Fusion Similarity Hashing (FSH).FSH explicitly embeds the graph based fusion similarity across modalities into a common Hamming space.

cross-modality similarity, since different similarities may be complementary to each other.

Recently, several works have been enhanced for multiple similarity measures[25, 32, 1], but little attention has been paid to preserve such similarity in a discrete Hamming space. In particular, it is shown that binary code learning by *fusion similarity* is more robust to noise compared with that by indirectly preserving intra- and inter-modal similarity respectively. However, it is not an easy task at all. The biggest concern lies in the efficiency issue in building the fusion model, *i.e.*, the fusion graph, which typically needs relaxation on the eigen decomposition of the graph Laplacian, resulting in significant performance degeneration with the growth of hash bits.

To address the above problems, we propose a novel cross-modality hashing method, termed Fusion Similarity Hashing (FSH), which makes the attempt towards directly preserving the fusion similarity from the multiple modalities to a common Hamming space. Such fusion similarity is robust to noise in capturing multi-modal relationship among instances. Different from the existing work of cross-modality hashing [12, 23], we argue that it is the fusion similarity, rather than the individual intra-modal similarity, that should be preserved in the common Hamming space. To that effect, an asymmetrical fusion graph is built, which simultaneously captures the intrinsic relations according to heterogeneous and homogenous data with a low storage cost afterwards.. After that, we design an efficient objective function to learn accurate binary codes and the corresponding hash functions in an alternating optimizing way. The whole framework of the proposed FSH approach is shown in Fig.1. FSH first builds the similarity matrix in each modality, and then combining them to construct a matrix that reflect the fusion similarity. According to such similarity, we use a asymmetric graph to learning the hash function, but it is hard to train the objective model due to discrete constraint. To handle such problem, we propose an alternating optimization algorithm, which also updates the fusion parameters, so as to find the optimal fusion graph to generate more discriminated hash codes.

We compare the proposed FSH approach against various state-of-the-art unsupervised cross-modality hashing methods [12, 21, 5, 10] on three large-scale benchmarks, *i.e.*, UCI Handwritten Digit , MIR-Flickr25K and NUS-WIDE. Our quantitative results demonstrate that FSH outperforms the existing unsupervised methods on standard benchmarks for four retrieval tasks.

2. Fusion Similarity Hashing

We first give notations used in the rest of this paper. Assume that $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ is the training set with n instances, where $o_i = (x_i^1, x_i^2, ..., x_i^M)$ is the i-th instance containing M feature vectors from M modalities respectively. $\mathbf{X}^m = \{x_1^m, x_2^m, ..., x_n^m\} \in \square^{d_m \times n}$ is defined as the feature matrix for the m-th modality, and $x_i^m \in \square^{d_m}$ is the i-th data of $\mathbf{X}^m$ with d_m dimension. We further denote $S_m(o_i, o_j) = \|x_i^m - x_j^m\|^2$ as the function to measure the Euclidean distance of o_i and o_j in the m-th modality. As mentioned, a graph matrix $\mathbf{G}$ is constructed to measure the fusion similarity among training instances, where $\mathbf{G}(i, j)$ indicates the affinity between instance o_i and o_j.

Given training data set $\mathcal{O}$, the proposed FSH aims to learn a set of hash functions $H^m(x^m) = \{h_1^m(x^m), ..., h_r^m(x^m)\}$ for the m-th modal data, and simultaneously learn the corresponding binary codes $\mathbf{B}^m = \{b_1^m, b_2^m, ..., b_n^m\} \in \{-1, 1\}^{r \times n}$ for the training data, where r is the code length. For the m-th modality data, its hash function can be written as:

$$h_k^m(x^m) = sgn\big(f_k^m(x^m)\big), \big(k = 1, 2, ..., r\big), \qquad (1)$$

where $sgn(\cdot)$ is the sign function, which returns 1 if $f_k^m(\cdot) > 0$ and -1 otherwise. $f_k^m(\cdot)$ is the linear or nonlinear transform function for data of the m-th modality. For simplicity, we define our hash function at the m-th modality as $H^m(x^m) = sgn(\mathbf{W}_m^T x^m)$.

Ideally, if the instances o_i and o_j are similar, the Hamming distance between their binary codes should be minimal, and vice versa. We do this by minimizing the quantization error between the fusion similarity matrix $\mathbf{G}$ and the Hamming similarity matrix $\mathbf{G}_H$, which can be written as:

$$min\|\mathbf{G} - \mathbf{G}_H\|_F^2, \qquad (2)$$

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

Therefore, the key issue falls in the construction quality of the fusion similarity matrix $\mathbf{G} \in \square^{n \times n}$. Inspired by the Neighbor Set Similarity (NSS) [1], it is straightforward to define our fusion similarity in the following: Given two instances with two modalities $o_i = \{x_i^m, x_i^t\}$ and $o_j = \{x_j^m, x_j^t\}$, the bi-modal NSS in the m-th modality

can be defined as:

$$\mathbf{S}_m\big(N_k^m(o_i), N_k^t(o_j)\big) = \frac{1}{k^2}\sum_{q\in N_k^m(o_i)}\sum_{y\in N_k^t(o_j)} S_m(o_q, o_y), \tag{3}$$

where $N_k^m(\cdot)$ returns the k-nearest neighbor index numbers according to the m-th modal similarity measure. And the fusion similarity $\mathbf{G}(i,j)$ across such two modalities can be defined via:

$$\begin{aligned}&\mathbf{G}(i,j) \\ &= mean\big\{\mathbf{S}_1\big(N_k^1(o_i), N_k^2(o_j)\big), \mathbf{S}_2\big(N_k^2(o_i), N_k^1(o_j)\big)\big\},\end{aligned} \tag{4}$$

It is quite intuitive to extend the above bi-modal fusion similarity to multi-modal case, *i.e.*

$$\begin{aligned}\mathbf{G}(i,j) = \frac{1}{M}\sum_{m=1}^{M}\big(\eta_m^\lambda \sum_{t\neq m}\mathbf{S}_m(N_k^m(o_i), N_k^t(o_j))\big) \\ s.t.\ \sum_{m=1}^{m}\eta_m = 1, 0\leq\eta_m\leq 1,\end{aligned} \tag{5}$$

where η_m refers to weight of the m-th modality, and λ controls the weight distribution across multiple modalities.

2.1. Sample-Importance Anchor Graph

However, with the increasing amount of data, the iteration of discovering the kNN set and computing NSS are both computationally intensive. To address this issue, inspired by the anchor graph based acceleration [18], we further propose an asymmetric similarity matrix, termed Sample-Importance Anchor Graph (SIAG), which jointly considers the diversity among the anchor points.

In particular, given an anchor set $\mathcal{L} = \{l_1, l_2, ..., l_p\}$, where $l_i = \{l_i^1, l_i^2, ..., l_i^M\}$ is the i-th anchor across M modalities, which are randomly sampled from the original data set $\mathcal{O}$. The key design here lies in the weighting of each l_i^m in the similarity matrix, for which the simplest way is to assign an identical weight by following the uniform distribution. However, due to the limitation of sampling and feature representation, anchor points are not uniform in the feature space.

To handle this, we use both probabilistic graph model and Markov Chain to predict the weight of each anchor as below. Firstly, we construct an undirected weighted graph $\mathbf{Z}^m \in \square^{p\times p}$ among anchors at the m-th modality, where $\mathbf{Z}_{i,j}^m = S_m(l_i^m, l_j^m)$. This adjacency matrix $\mathbf{Z}^m$ can be interpreted in a probabilistic way [16]. According to the Markov networks, the statable transition network $\hat{\mathbf{Z}}$ can be achieved by Markov random walks. Then, according to the new graph matrix $\hat{\mathbf{Z}}^m$, the weighted vector $\alpha^m = \{\alpha_1^m, ..., \alpha_p^m\}$ is learned by using the Markov Chain prediction, where α_i^m is the weight parameter of the i-th anchor in the m-th modality.

By far, the Sample-Importance Anchor Graph (SIAG) is built. We then use the SIAG to rewrite the fusion similarity:

$$\begin{aligned}\mathbf{G}(o_i,o_j) &= \frac{1}{M}\sum_{m=1}^{M}\eta_m^\lambda \mathbf{L}_m(o_i,o_j), \\ \mathbf{L}_m(o_i,o_j) &= \frac{1}{k^2}\sum_{t\neq m}\hat{\mathbf{S}}_m\big(N_{k,l}^m(o_i), N_{k,l}^t(o_j)\big), \\ \hat{\mathbf{S}}_m\big(N_{k,l}^m(o_i), N_{k,l}^t(o_j)\big) &= \frac{1}{k^2}\sum_q\sum_y \alpha_q^m\alpha_y^m S_m(l_q, l_y),\end{aligned} \tag{6}$$

where $q\in N_{k,l}^m(o_i)$, $y\in N_{k,l}^t(o_j)$, and $N_{k,l}^m(\cdot)$ returns the index numbers of k-nearest anchors in the m-th modality.

2.2. The Proposed FSH Scheme

In terms of solving the minimization in Eq.2, the main objective is to learn a set of binary codes $\{\mathbf{B}^1, \mathbf{B}^2, ..., \mathbf{B}^M\}$ for the M modalities respectively, whose inner product is expected to approximate the similarity matrix $\mathbf{G}$. Inspired by [17, 22], we formulate the following problem of cross-modality similarity preserving with binary codes $\mathbf{B}^m$ and hybrid graph $\mathbf{S_F}$ as,

$$\min_{\mathbf{B}^m}\sum_{m=1,t\neq m}^{M}\|\mathbf{B}^{mT}\mathbf{B}^m - \mathbf{G}\|_F^2 + \gamma\|\mathbf{B}^{mT}\mathbf{B}^t - \mathbf{G}\|_F^2, \tag{7}$$

where γ is a balance parameter.

Intuitively, we learn the m-th modality hash function $H^m(\mathbf{X}^m)$ by minimizing the error term between the linear hash function in Eq.1, constrained by the corresponding binary code $\mathbf{B}^m$ by $\|\mathbf{B}^m - H^m(\mathbf{X}^m)\|_F^2$. Such hash function learning can be easily integrated into the overall cross-modality similarity persevering, which is rewritten as:

$$\begin{aligned}\min_{\mathbf{B}^m,\mathbf{W}^t}&\textstyle\sum_{t\neq m}\|\mathbf{B}^{mT}\mathbf{B}^m - \mathbf{G}\|_F^2 + \gamma\|\mathbf{B}^{mT}\mathbf{B}^t - \mathbf{G}\|_F^2 \\ &+\mu\textstyle\sum_{m=1}^{M}\|\mathbf{B}^m - H^m(\mathbf{X}^m)\|_F^2 \\ s.t.\ &\mathbf{B}^m\in\{-1,1\}^{r\times n},\ \textstyle\sum_{m=1}^{m}\eta_m = 1,\ 0\leq\eta_m\leq 1,\end{aligned} \tag{8}$$

where μ is a tradeoff parameter to control the weights between minimizing the binary quantization and preserving the fusion similarity.

Ideally, binary codes from different modalities of the same instance should be set as identical as possible, which is similar to the previous Cross-Modality Hashing, *i.e.*, CMFH [5] and SMFH [15]. We further set the constraints $\mathbf{B} = \mathbf{B}^1 = \mathbf{B}^2$ and $\mathbf{B}\mathbf{B}^T = \mathbf{I}$ in Eq.8 with which the first and second terms in Eq.8 can be integrated, and the balance parameter λ can be neglected. Correspondingly, we rewrite the overall FSH as:

$$\begin{aligned}\min_{\mathbf{B},\mathbf{W}^t}&\|\mathbf{B}^T\mathbf{B} - \mathbf{G}\|_F^2 + \mu\textstyle\sum_{m=1}^{M}\|\mathbf{B} - H^m(\mathbf{X}^m)\|_F^2 \\ s.t.\ &\mathbf{B}\in\{-1,1\}^{r\times n},\ \textstyle\sum_{m=1}^{m}\eta_m = 1,\ 0\leq\eta_m\leq 1,\end{aligned} \tag{9}$$

where $Tr(\cdot)$ is the trace of the matrix.

However, the scale of the matrix $\mathbf{G}$ is extremely large, which needs huge storage cost and makes Eq. 9 hard to optimize on a large training set. Therefore, to handle this problem, the symmetric fusion similarity matrix $\mathbf{G}$ can be approximated by Cholesky decomposition $\mathbf{G} = \mathbf{U}\mathbf{U}^T$, where $\mathbf{U} \in \mathbf{R}^{n\times p}$, and $\mathbf{U}$ can be represented as the low-rank approximated to the high-dimensional matrix $\mathbf{G}$. Although such decomposition makes the storage efficiency, it is still too complexity to calculate such decomposition. Similar to anchor graph [18], we rewrite the proposed SIAG in Eq.6 to approximate the matrix $\mathbf{U}$,

$$\mathbf{U}(i,j) = \hat{\mathbf{G}}(i,j) = \mathbf{G}(o_i, l_j) \in \mathbf{R}^{n\times p}. \tag{10}$$

According to Theorem 2 in [6], $Tr(\mathbf{G}^T\mathbf{G}) \leq \beta Tr(\mathbf{G})$, when $\beta \geq \Lambda_1$, and Λ_1 is the largest eigen-value of matrix $\mathbf{G}$. The first item in Eq.9 can be approximated rewritten as:

$$\min_B \beta Tr(\mathbf{G}) - 2Tr(\mathbf{B}\mathbf{G}\mathbf{B}^T).$$

Then, we assume $\mathbf{B_s} = sgn(\mathbf{B}\hat{\mathbf{G}}) \in \{1,-1\}^{r\times p}$ to be the binary anchors, and $\mathbf{U}$ to be the affinity matrix that measures the fusion similarity between data points and anchors' binary codes $\mathbf{B_s}$. To learn the binary codes, the overall objective function can be written as:

$$\begin{aligned} \min_{\mathbf{B},H^m} & \beta Tr(\hat{\mathbf{G}}^T\hat{\mathbf{G}}) - 2Tr(\mathbf{B}\hat{\mathbf{G}}{\mathbf{B_s}}^T) \\ & + \mu\textstyle\sum_{m=1}^{M} \|\mathbf{B} - H^m(\mathbf{X}^m)\|_F^2 \\ & s.t.\ \textstyle\sum_{m=1}^{m} \eta_m = 1,\ 0 \leq \eta_m \leq 1. \end{aligned} \tag{11}$$

2.3. Optimization

Directly minimizing the objective function in Eq. 11 is intractable due to the discrete constraint of hash functions. The formulation is non-convex with respect to $\mathbf{B}$, $\mathbf{B_s}$, $\mathbf{W_m}$, $\hat{\mathbf{G}}$, and η_m jointly. This is further handled by using an alternating optimization, *i.e.*, updating one variable with fixing the rest three until convergence.

(1) Fix $\mathbf{B_s}$, $\mathbf{W_m}$, $\hat{\mathbf{G}}$ and η_m, then update $\mathbf{B}$. The corresponding sub-problem is:

$$\begin{aligned} \min_{\mathbf{B}} & -2Tr(\mathbf{B}\hat{\mathbf{G}}{\mathbf{B_s}}^T) + \mu\textstyle\sum_{m=1}^{M} \|\mathbf{B} - \mathbf{W}_m^T\mathbf{X}^m\|_F^2 \\ & s.t.\ \mathbf{B} \in \{-1,1\}^{r\times n}, \mathbf{B_s} \in \{-1,1\}^{r\times p}, . \end{aligned} \tag{12}$$

Solving Eq.12 is still not convenient. To this end, we fix the η_m and $\hat{\mathbf{G}}$ at the same time, then it can be expended into:

$$\begin{aligned} O_1(\mathbf{B}) = & -2Tr(\mathbf{B}\hat{\mathbf{G}}{\mathbf{B_s}}^T) + \mu M Tr(\mathbf{B}^T\mathbf{B}) \\ & - 2\mu\sum_{m=1}^{M} Tr(\mathbf{B}{\mathbf{X}^m}^T\mathbf{W}_m), \end{aligned} \tag{13}$$

In this way, the gradient of Eq.13 is given by:

$$\frac{\partial O_1}{\partial \mathbf{B}} = -\hat{\mathbf{G}}{\mathbf{B_s}}^T + \mu M\mathbf{B}^T - \mu\sum_{m=1}^{M} {\mathbf{X}^m}^T\mathbf{W}_m. \tag{14}$$

Let $\frac{\partial O_1}{\partial \mathbf{B}} = 0$, this sub-problem can be solved by the following updating rule:

$$\mathbf{B}^T = sgn\big(\frac{1}{\mu M}(\hat{\mathbf{G}}{\mathbf{B_s}}^T + \mu\sum_{m=1}^{M} {\mathbf{X}^m}^T\mathbf{W}_m)\big). \tag{15}$$

(2) Fix $\mathbf{B}$, $\mathbf{W_m}$, $\hat{\mathbf{G}}$, and η_m, then update $\mathbf{B_s}$. When fixing $\mathbf{B}$ and $\mathbf{W_m}$, the updating of $\mathbf{B_s}$ can be referred to:

$$\max_{\mathbf{B}} Tr(\mathbf{B}\hat{\mathbf{G}}{\mathbf{B_s}}^T). \tag{16}$$

With the same scheme to handle sup-problem (1), this sub-problem is solved as follow:

$$\mathbf{B_s} = sgn(\mathbf{B}\hat{\mathbf{G}}), \tag{17}$$

which is similar to the before assumption.

(3) Fix $\mathbf{B}$, $\mathbf{B_s}$, $\hat{\mathbf{G}}$, and η_m, then update $\mathbf{W_m}$. This sub-problem finds the best projection coefficient $\mathbf{W_m}$ by minimizing $\|\mathbf{B} - \mathbf{W}_m^T\mathbf{X}^m\|_F^2$ with the traditional linear regression. Therefore, we update $\mathbf{W_m}$ with other variables fixed:

$$\mathbf{W_m} = (\mathbf{X}^m{\mathbf{X}^m}^T + \sigma\mathbf{I})^{-1}\mathbf{X}^m\mathbf{B}, \tag{18}$$

where β is the regularization parameter, and $\mathbf{I}$ is the identity matrix.

(4) Fix $\mathbf{B}$, $\mathbf{B_s}$, $\hat{\mathbf{G}}$, and $\mathbf{W_m}$, then update η_m. The last sub-problem is to minimize Eq.11 with respect to the weight η_m. We use the Lagrange Multiplier with constraint $\sum_{m=1}^{M} \eta_m = 1$ as:

$$O_2(\eta_m) = \beta Tr(\hat{\mathbf{G}}^T\hat{\mathbf{G}}) - 2Tr(\mathbf{B}\hat{\mathbf{G}}{\mathbf{B_s}}^T) - \alpha(\sum_{m=1}^{M} \eta_m - 1). \tag{19}$$

However, this problem is not a convex. We then omit the first constant item and replace the third item with matrix $\mathbf{A}^* = \sum_{m=1}^{M} \eta_m^\lambda \mathbf{A_m}$, where $\mathbf{A_m}(i,j) = S_m(l_i^m, l_j^m)$ is the fusion similarity at the m-th modality among anchor points. It can be rewritten as follows:

$$\hat{O}_2(\eta_m) = \sum_{m=1}^{M} \eta_m^\lambda P_m - \alpha(\sum_{m=1}^{M} \eta_m - 1), \tag{20}$$

where $P_m = Tr\big(\beta\mathbf{A_m} - 2{\mathbf{B_s}}^T\mathbf{B}\mathbf{L_m}(\mathcal{O},\mathcal{L})\big)$. Then this sub-problem can be solved by the traditional weighted learning scheme, which has been widely used in previous works [3, 13, 19]. The updating rule of this sub-problem is:

$$\eta_m = (\lambda P_m)^{\frac{1}{1-\lambda}} \Big/ \sum_{m=1}^{M} (\lambda P_m)^{\frac{1}{1-\lambda}}. \tag{21}$$

(4) Fix $\mathbf{B}$, $\mathbf{B_s}$, $\mathbf{W_m}$ and η_m, then update $\hat{\mathbf{G}}$. The new fusion similarity matrix $\hat{\mathbf{G}}$ is updated upon the definition in Eq.6, which combines each modal asymmetric similarity to construct the matrix that reflect more accurate fusion similarity. We summarize the overall procedure of the proposed FSH scheme in Algorithm 1, which leverages bi-modal data as an example.

2.4. Convergence Proof

We prove the convergence of Eq.20 with the alternative method by the following lamma.

Lamma 1. For the asymmetric similarity matrix $\hat{\mathbf{G}} \in \square^{n\times p}$, its multiplication $\hat{\mathbf{G}}^T\hat{\mathbf{G}} \in \square^{p\times p}$ approximately estimates the fusion similarity of anchor points, which is defined as $\mathbf{A}^* = \sum_{m=1}^{M} \mathbf{A_m}$.

Proof. The average similarity between two anchor points l_i and l_j is denoted as $\mathbf{A}^*_{ij} = P(l_j|l_i)$. Therefore the average similarity by multi-modality anchor points can be regarded as a two-step transition probability through data at each modality x^m_k:

$$\mathbf{A}^*_{ij} = P(l_j|l_i) = \sum_{m=1}^{M}\sum_{k=1}^{n} p(l^m_j|x^m_k)p(x^m_k|l^m_i), \qquad (22)$$

where $p(l^m_j|x^m_k) = S_m(l^m_j, x^m_k)/\sum_k S_m(l^m_j, x^m_k)$. Then we define $\mathbf{Q} = \hat{\mathbf{G}}^T\hat{\mathbf{G}}$, and let $\eta^\lambda_m = v_m, 0 \le v_m \le 1$, leading to,

$$\begin{aligned}
\mathbf{Q}(i,j) &= \sum_{k=1}^{n} \hat{\mathbf{G}}(k,i)\hat{\mathbf{G}}(k,j) \\
&= \sum_k \Big\{ \sum_m v_m p(x^m_k|l^m_i) \sum_m v_m p(x^m_k|l^m_j) \Big\} \\
&= \sum_k \Big\{ \sum_m v^2_m p(x^m_k|l^m_i)p(x^m_k|l^m_j) \\
&+ \sum_{t\neq m} v_m v_t p(x^m_k|l^m_i)p(x^t_k|l^t_j) \Big\}.
\end{aligned} \qquad (23)$$

Obviously, taking the condition $v^2_m \le v_m$ and $0 \le v_m v_t \le 1$ into Eq.24, the following conclusion can be got,

$$\begin{aligned}
&\sum_k \sum_m v_m p(x^m_k|l^m_i)p(x^m_k|l^m_j) \le \mathbf{Q}(ij) \\
&\le \sum_k \sum_m v_m p(x^m_k|l^m_i)p(x^m_k|l^m_j) + Const.
\end{aligned} \qquad (24)$$

Summing Eq.22 and Eq.24 in the two sides, we arrive at $\mathbf{A}^* \le \mathbf{Q} \le \mathbf{A}^* + Const$. Due to the normalization of similarity graph, the $Const$ parameter is moderate, which makes a small the estimation error. Thus the Eq.20 will be converged with an alternative method. □

Algorithm 1 Fusion Similarity Hashing (FSH)

Input: Training data set $\mathcal{O}$ with two modalities $\mathbf{X}^1$ and $\mathbf{X}^2$, the number of anchor points *m*, the number of hash bits *r*, and the parameters k and μ.

Output: The hash codes $\mathbf{B}$ for training instances $\mathcal{O}$ and the projection coefficient matrix $\mathbf{W}^t$.

```
Initialize W^1, W^2 by CCA method [8] .
Uniformly and randomly select p sample pairs from training instances as the anchors L.
Initialize hash codes B and B_s by CVH [12].
Initialize η_m = 1/M.
Construct graph Ĝ and A_m for m-th modality.
repeat
    Fix W^t, B_s, Ĝ, and η_m, update B by Eq. 15;
    Fix B, W^t, Ĝ, and η_m, update B_s by Eq. 17;
    Fix B, B_s, Ĝ, and η_m, update W^t by Eq. 18;
    Fix B, B_s, Ĝ, and W^t, update η_m by Eq. 21;
    Fix B, B_s, W^t, and η_m, update Ĝ by by Eq. 10;
until convergence or reaching the maximum iteration.
```

3. Experiments

In this section, we conduct a serial of quantitative experiments to validate the proposed FSH algorithm on three widely-used benchmarks, *i.e.*, **UCI Handwritten Digit**[1], **MIR-Flickr25K**[2], and **NUS-WIDE**[3].

3.1. Competing Methods

We evaluate the cross-modality retrieval task via: (1) the image-modality to tag-modality side, termed **Task 1**, (2) the tag-modality to image-modality side, termed **Task 2**, (3) the image-modality to image-modality side, termed **Task 3**, and (4) the tag-modality to tag-modality side, termed **Task 4**. In the above tasks, the proposed **FSH** is compared against four state-of-the-art unsupervised methods, *i.e.*, Cross-View Hashing (**CVH**) [12], Predictable Dual-view Hashing (**PDH**) [21], Collective Matrix Factorization Hashing (**CMFH**) [5], and Alternating Co-Quantization (**ACQ**) [10]. Except these, we further compare the **FSH** with a simple fusion graph construction, which just fuse the anchor graph in each modality, and we refer this as **FSH-S** which is commonly a strong baseline for both two cross-modality retrieval tasks[4]. All the source codes of the rest methods are available publicly, and we directly adopt the original parameter settings described in their papers. All our experiments were run on a workstation with a 3.6GHz Intel Core I7 − 4790 CPU and 16G RAM.

[1] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[2] http://www.cs.toronto.edu/nitish/multimodal/
[3] http://lms.comp.nus.edu/research/NUSWIDE.htm
[4] It means that $\mathbf{L}_m(i,j) = \frac{1}{2}\sum_i \sum_j S_m(o_i, l_j)$.

3.2. Datasets

The **UCI Handwritten Digit** dataset consists of multi-modal features of handwritten numerals ($0-9$), which are extracted from a collection of Dutch utility maps. It contains 10 categories, each of which has 200 patterns. Following the setting of [9], select 76 Fourier coefficients of the character shapes as one modal features, and 64 Karhunen-Love coefficients as the other modal features, $1,500$ images are randomly sampled as the training set, and the remaining as query.

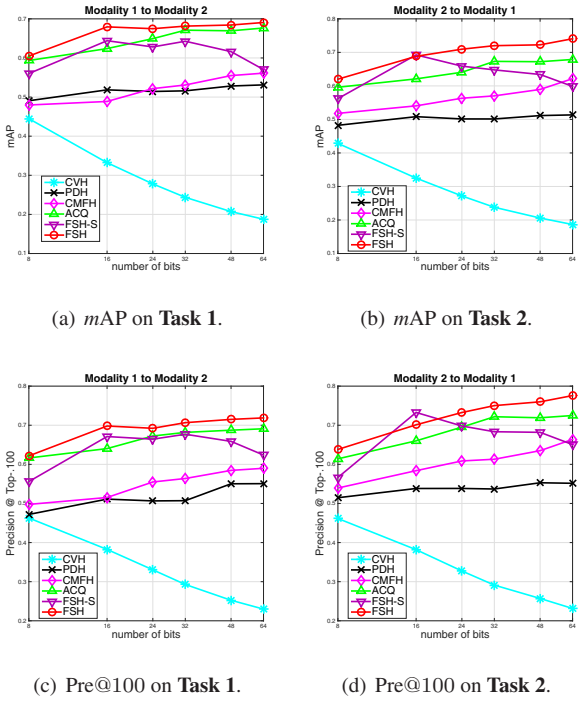
The **MIR-Flickr25K** dataset is collected from Flickr website, which contains $25,000$ images together with 24 provided unique labels. The image for each instance is described by the 150 dimension Edge histogram descriptor which mainly encodes surface texture. And the 500-dimensional feature vector, derived from PCA on its binary tagging vectors, is used for text representation. Follow the setting of [5], the above descriptors are defined as two modalities for our cross-modality retrieval task. We take out 5% of the dataset as the query set, and the remaining as the training set.

The **NUS-WIDE** dataset is a real-world web image dataset crawled from Flickr, which contains $296,648$ images with associated tags. Each image-tag pair is annotated with one or more labels from 81 concepts. We select $186,577$ labeled image-tag pairs from the whole dataset according to the top 10 largest concepts, as adopted in [30, 33]. 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 $2,000$ image-tag pair points from this database as the query set, and the remaining as the training data set. And we analyze the parameters of the proposed FSH by fixing the code length to 64. The analysis is done by varying one value while fixing the others.

3.3. Evaluation Protocols and Parameter Settings

The quantitative performance is evaluated by mean Average Precision (*m*AP). *m*AP is the mean of Average precision (AP) over all queries, which jointly considers search accuracy and rankings. Given a query and a list of retrieval results, AP is defined as $\sum_{i=1}^{n} p(i)\delta(i)$, where $p(i)$ denotes the precision of the top i retrieved images, and $\delta(i) = 1$ iff the r-th retrieved image is the true neighbor of the query, otherwise $\delta(i) = 0$. We also consider other three evaluation protocols, *i.e.*, Precision at top-100 positions (termed Pre@100), and Precision curves at top-K (termed Rec@K).

In our experiments, the parameter μ is a trade-off parameter, which is set as 300 on three datasets. The regularization parameter σ in Eq.18 is set to be a small number 0.0001 in all experiments. The weight controller parameter λ is set with 5-fold cross validation using the training data. The parameter k controls the effectiveness of the hybrid similarity, which is tuned in the next subsection, and set as 10 in all our experiments. For training efficiency, the number of anchors is set to 100 for all the datasets.



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

(c) Pre@100 on **Task 1**. (d) Pre@100 on **Task 2**.

Figure 2. The *m*AP curves and Precision@100 curves of all the algorithms on **UCI Handwritten Digit**. The *m*AP evaluation is shown in the first row, and Pre@100 are shown in the bottom. Best view in color.

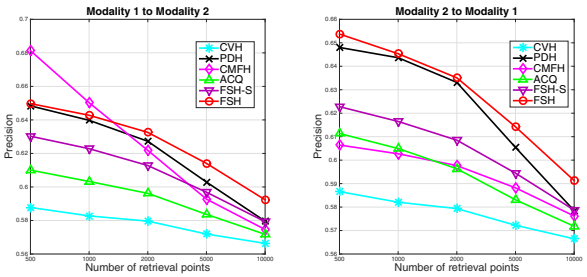
3.4. Quantitative Results

For the UCI Handwritten Digit dataset, our quantitative results are shown in Fig.2, Tab.1 and Tab.2. It demonstrates that our proposed FSH has achieved superior performance on the UCI benchmark for all the four retrieval tasks. The first two subfigures in Fig.2 show the remarkable *m*AP result in the cross-modality retrieval task. Comparing to the second best scheme **ACQ**, the proposed FSH has achieved about 3.4% improvement for the first task and about 8.1% for the second.[5] The Pre@100 for Task 1 and Task 2 is shown in Fig.2 (c) and (d). It is worth noting that, for both tasks, our FSH shows advantage on precision with all hash bits, which is mainly due to the fact that more accurate similarity got from fusion similarity with SIAG can find more optimal binary codes from multi-modality. We further compare the simple fusion way, named **FSH-S**, with the baselines [12, 21, 5, 10]. The results in Fig.2 and Tab.1 show that the simple fusion model can also achieve second best

[5]The percentage of *m*AP growth is obtained by the means of improvement on all hash bits.

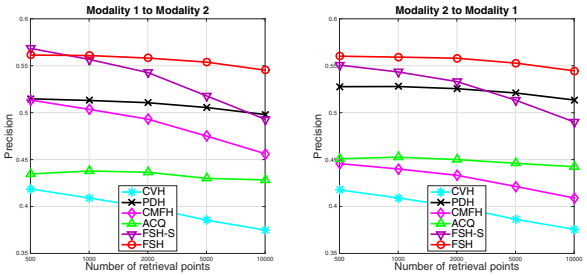
Table 1. The *m*AP and Precision Comparison Using Hamming Ranking on Two Benchmark with Different Hash Bits

Task	Methods	MIR-Flickr25K					NUS-WIDE				
		*m*AP			Pre@100		*m*AP			Pre@100	
		16	32	64	32	64	16	32	64	32	64
Task1	CVH	0.5819	0.5756	0.5710	0.6119	0.5972	0.3811	0.3685	0.3574	0.4749	0.4387
	PDH	0.5981	0.6026	0.6043	0.6517	0.6598	0.4658	0.4747	0.4780	0.4989	0.5125
	CMFH	0.5839	0.5854	0.5857	**0.7019**	**0.7225**	0.3723	0.3781	0.3799	0.5064	0.5309
	ACQ	0.5871	0.5857	0.5823	0.6223	0.6231	0.4247	0.4435	0.4328	0.4442	0.4309
	FSH-S	**0.6090**	0.5969	0.5930	0.6401	0.6429	0.4996	0.4610	0.4556	**0.5698**	**0.5876**
	FSH	0.5968	**0.6189**	**0.6195**	0.6555	0.6629	**0.5059**	**0.5063**	**0.5171**	0.5297	0.5616
Task2	CVH	0.5803	0.5750	0.5708	0.6086	0.5923	0.3768	0.3652	0.3555	0.4690	0.4321
	PDH	0.5941	0.5976	0.5997	0.6508	0.6548	0.4458	0.4519	0.4552	0.5133	0.5284
	CMFH	0.5960	0.5942	0.5960	0.6230	0.6192	0.3957	0.4036	0.4105	0.4517	0.4595
	ACQ	0.5857	0.5829	0.5815	0.6292	0.6258	0.4134	0.4273	0.4200	0.4573	0.4887
	FSH-S	**0.6036**	0.5944	0.5923	0.6458	0.6356	0.4776	0.4460	0.4423	**0.5521**	0.5597
	FSH	0.5924	**0.6128**	**0.6091**	**0.6657**	**0.6689**	**0.4790**	**0.4810**	**0.4965**	0.5388	**0.5685**



(a) Pre@K on **Task 1**. (b) Pre@K on **Task 2**.

(1) MIR-Flickr25K Dataset



(c) Pre@K on **Task 1**. (d) Pre@K on **Task 2**.

(2) NUS-WIDE Dataset

Figure 3. The Precision curves and Recall curves of all the algorithms on two benchmark when hash bit is 64.

performance, which demonstrate that simple or complexity fusion similarity should be preserved in binary codes. However, FSH is overall better than FSH-S. As shown in [1], NSS is robust to noise with different similarity measures, and such similar scheme in fusion construction makes the proposed FSH more robust for cross-modality retrieval task.

As shown in Tab.2, we conduct the experiments on Task 3 and Task 4, which are both single-modality retrieval tasks. For both tasks, we replace CVH and CMFH with two classical single-modality hashing, *i.e.*, Iterative Quantization (ITQ) [7] and Scalable Graph Hashing (SGH) [11]. Intuitively, combing multiple features improves the search performance. However, from Tab.2, the *m*AP scores of the cross-modality hashing schemes, except FSH, are not lower than ITQ in Task 4. To explain, traditional single-modality ITQ and SGH have higher *m*AP scores in Task 4 than that in Task 3, reflecting that the expression power of the 64-Dim Karhunen-Love coefficients is much better than that of the 76-Dim Fourier coefficient features. Such diversity leads to noise in similarity measurement when learning cross-modality binary codes, which is ignored in the previous works [12, 23] , the cross-modality hashing always finds a common Hamming subspace or real-valued subspace to approximate the optimal solution. For the proposed FSH, the performance is not only better in cross-modality retrieval, but also very competitive in single-modality retrieval.

For the MIR-Flickr25K benchmark, as shown in Tab.1, Tab.2 and the first row of Fig.3, FSH has achieved the overall best performance for all the four retrieval tasks. The *m*AP and Pre@100 results on this benchmark are reported in Tab.1 and Tab.2, under the settings of $16, 32$, and 64 bits respectively. FSH has achieved remarkable *m*AP and precision scores. The Precision and Recall curves are shown in the first row of Fig.3. As the same result in Pre@100, FSH has achieved comparable and state-of-the-art performance comparing to all the baselines [12, 21, 5, 10]. Although CMFH achieves better Pre@100 in the Task 1, FSH has the better result in *m*AP with all the hash bits from Tab.1. To explain these problems, the *m*AP is the global ranking evaluation and the Pre@100 is the local ranking evaluation. It shows that CMFH has better performance in terms of searching truth neighbors quickly, but FSH achieves better *m*AP in terms of finding more true candidates. For Pre@100, FSH is still the second one on Task 1 and the first on Task 2, which claims that the overall performance of the FSH is better on the MIR-Flickr25K.

Table 2. The *m*AP Results Using Hamming Ranking on Three Benchmark with Different Hash Bits on Task 3 and Task 4.

		UCI Handwritten Digit			MIR-Flickr25K			NUS-WIDE		
		16	32	64	16	32	64	16	32	64
Task 3	ITQ	0.5325	0.5491	0.5475	0.5685	0.5695	0.5705	0.4692	0.4585	0.4401
	SGH	0.4380	0.4249	0.4425	0.5661	0.5676	0.5692	0.4367	0.4127	0.4010
	PDH	0.5007	0.4991	0.5133	0.6016	0.6073	0.6093	0.4671	0.4760	0.4782
	ACQ	0.6100	0.6616	0.6637	0.5915	0.5895	0.5876	0.4191	0.4393	0.4289
	FSH	**0.6380**	**0.6626**	**0.6669**	**0.5998**	**0.6233**	**0.6282**	**0.5015**	**0.4965**	**0.5140**
Task 4	ITQ	0.6322	0.6397	0.6554	0.5652	0.5665	0.5671	0.3774	0.3769	0.3740
	SGH	0.5497	0.5972	0.6244	0.5602	0.5580	0.5562	0.3730	0.3686	0.3706
	PDH	0.4841	0.5344	0.5429	0.6027	0.6073	0.6094	0.4431	0.4476	0.4497
	ACQ	0.5892	0.6527	0.6498	0.5909	0.5881	0.5860	0.4075	0.4239	0.4155
	FSH	**0.6364**	**0.6607**	**0.6684**	**0.5991**	**0.6201**	**0.6152**	**0.4760**	**0.4754**	**0.4936**

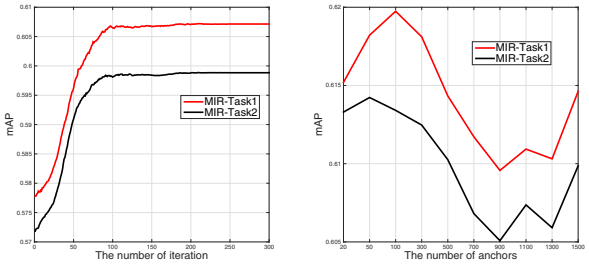
On the large-scale NUS-WIDE dataset, as shown in Tab.1, Tab.2 and Fig.3, FSH still achieves the highest search accuracy. It is noted that FSH has significant improvement in all the four retrieval tasks. When comparing with the second highest method, the improvement of *m*AP score is 8.2% in Task 1 with 64 hash bit, and 9.1% in Task 2 with 64 bits. This demonstrates that the fusion similarity has advantageous to produce more distinguished binary code on the modality with weak expression power, which subsequently enhances the performance of single-modality retrieval.

3.5. Parameter Analysis

In this subsection, we analyze the parameters of the proposed FSH by fixing the code length to 64. The analysis is done by varying one value while fixing the others.

As shown in Fig.4 (a), the *m*AP of our methods does not change significantly with more than 100 iterations, which also holds for the other two datasets. Therefore in our experiments, we fix the iteration number to 100 for quickly alternating optimization.

We further plot the *m*AP score of Hamming ranking with the increasing number of landmarks ($20 \leq p \leq 1,500$) in Fig.4 (b). We observe that *m*AP decreases with the growing number of landmarks. It is shown that a large size of anchors with small parameter k will bring more noise to the fusion graph, which decreases the performance of the proposed FSH. As a result, the relation between the size of anchors and parameter k is empirically calculated by $\#anchor = 10 \times k$, which can achieve satisfactory results in both datasets. As a conclusion, for the proposed FSH, asymmetric fusion graph with little anchor points can enhance the performance of cross-modality retrieval, which solves the large-scale problem of binary code learning efficiently.



(a) *m*AP vs. Iteration. (b) *m*AP vs. #anchors.

Figure 4. The parameter analysis. (The two are conducted on MIR-Flickr25K dataset.)

4. Conclusion

In this paper, we propose a novel hashing method termed Fusion Similarity Hashing (FSH) for cross-modality retrieval. The core idea is to directly preserve the fusion similarity into the Hamming space to explicitly capture their heterogeneous correlation in retrieval. To this end, a fusion graph is constructed to define the similarity among multi-modality instances. Then a graph hashing framework is proposed with alternating optimization, which learns consistent binary codes and the hash functions for each modality. Asymmetric discrete optimization is further used to train the model on large-scale data set. In this framework, combining neighbor set similarity with sample important anchor graph can be embedded to the fusion graph matrix, leading to the learning of more discriminative binary codes. Extensive experiments conducted on UCI Handwritten Digit, MIR-Flickr25K and NUS-WIDE benchmarks demonstrated the superior performance of FSH over several representative and state-of-the-art unsupervised cross-modality hashing methods.

5. Acknowledgement

This work is supported by the National Key Technology R&D Program (No. 2016YFB1001503), the Nature Science Foundation of China (No. 61422210, No. 61373076, No. 61402388, and No. 61572410).

References

[1] X. Bai, S. Bai, and X. Wang. Beyond diffusion process: Neighbor set similarity for fast re-ranking. *Information Sciences*, 325:342 – 354, 2015.

[2] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Proceeding of CVPR*, pages 3594 – 3601, 2010.

[3] X. Cai, F. Nie, W. Cai, and H. Huang. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *Proceeding of ICCV*, pages 1737–1744, 2013.

[4] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE TPAMI*, 36(3):521–535, 2014.

[5] G. Ding, Y. Guo, J. Zhou, and Y. Gao. Large-scale cross-modality search via collective matrix factorization hashing. *IEEE TIP*, 25(11):5427–5440, 2016.

[6] Y. Fang, K. A. Loparo, and X. Feng. Inequalities for the trace of matrix product. *IEEE TAC*, 39(12):2489–2490, 1994.

[7] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proceeding of CVPR*, pages 817–824, 2011.

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[9] R. He, M. Zhang, L. Wang, and Y. Ji. Cross-modal subspace learning via pairwise constraints. *IEEE TIP*, 24(12):5543–5556, 2014.

[10] G. Irie, H. Arai, and Y. Taniguchi. Alternating co-quantization for cross-modal hashing. In *Proceedings of ICCV*, pages 1886–1894, 2015.

[11] Q.-Y. Jiang and W.-J. Li. Scalable graph hashing with feature transformation. In *Proceedings of IJCAI*, 2015.

[12] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In *Proceeding of IJCAI*, pages 1360–1365, 2011.

[13] Y. Li, F. Nie, H. Huang, and J. Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceeding of AAAI*, pages 2750–2756, 2015.

[14] Z. Lin, G. Ding, M. Hu, and J. Wang. Semantics-preserving hashing for cross-view retrieval. In *Proceedings of CVPR*, pages 3864–3872, 2015.

[15] H. Liu, R. Ji, Y. Wu, and G. Hua. Supervised matrix factorization for cross-modality hashing. In *Proceeding of IJCAI*, pages 1767–1773, 2016.

[16] W. Liu, J. He, and S.-F. Chang. Large graph construction for scalable semi-supervised learning. In *Proceedings of ICML*, pages 679–686, 2010.

[17] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *Proceedings of CVPR*, pages 2074–2081, 2012.

[18] W. Liu, J. Wang, S. Kumar, and S.-F. Chang. Hashing with graphs. In *Proceedings of ICML*, pages 1–8, 2011.

[19] F. Nie, J. Li, and X. Li. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *Proceeding of IJCAI*, pages 1881–1887, 2016.

[20] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of MM*, pages 251–260, 2010.

[21] M. Rastegari, J. Choi, S. Fakhraei, D. Hal, and L. Davis. Predictable dual-view hashing. In *Proceedings of ICML*, pages 1328–1336.

[22] F. Shen, W. Liu, S. Zhang, Y. Yang, and H. Tao Shen. Learning binary codes for maximum inner product search. In *Proceedings of ICCV*, pages 4148–4156, 2015.

[23] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of SIGMOD*, pages 785–796, 2013.

[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of CVPR*, pages 3156–3164, 2015.

[25] B. Wang, J. Jiang, W. Wang, Z.-H. Zhou, and Z. Tu. Unsupervised metric fusion by cross diffusion. In *Proceeding of CVPR*, pages 2997–3004, 2012.

[26] Y. Wei, Y. Song, Y. Zhen, B. Liu, and Q. Yang. Scalable heterogeneous translated hashing. In *Proceedings of SIGKDD*, pages 791–800, 2014.

[27] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Proceeding of NIPS*, pages 1753–1760, 2009.

[28] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang. Quantized correlation hashing for fast cross-modal search. In *Proceedings of IJCAI*, 2015.

[29] T. Yao, T. Mei, and C.-W. Ngo. Learning query and image similarities with ranking canonical correlation analysis. In *Proceedings of ICCV*, pages 28–36, 2015.

[30] D. Zhang and W.-J. Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *Proceeding of AAAI*, pages 2177–2183, 2014.

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

[32] L. Zheng, S. Wang, Z. Liu, and Q. Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *Proceeding of CVPR*, pages 4321–4328.

[33] J. Zhou, G. Ding, and Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In *Proceedings of SIGIR*, pages 415–424, 2014.

[34] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of MM*, pages 143–152, 2013.