Online Cross-Modal Scene Retrieval by Binary Representation and Semantic Graph

Mengshi Qi
State Key Laboratroy of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University
Beijing, China 100191

Yunhong Wang
State Key Laboratroy of Virtual Reality Technology and Systems
School of Computer Science and Engineering, Beihang University
Beijing, China 100191

Annan Li*
Beijing Advanced Innovation Center for Big Data and Brain Computing,
Beihang University
Beijing, China 100191

ABSTRACT

In recent years, cross-modal scene retrieval has attracted more attention. However, most existing approaches neglect the semantic relationship between objects in a scene together with the embedded spatial layouts. Moreover, these methods mostly apply the batch learning strategy, which is not suitable for processing streaming data. To address the aforementioned problems, we propose a new framework for online cross-modal scene retrieval based on binary representations and semantic graph. Specially, we adopt the cross-modal hashing based on the quantization loss of different modalities. By introducing the semantic graph, we are able to extract wealthy semantics and measure their correlation across different modalities. Further more, we propose a two-step optimization procedure based on stochastic gradient descent for online update. Experimental results on four datasets show the superiority of our approach over the state-of-the-art.

CCS CONCEPTS

•Information systems →Image search;

KEYWORDS

scene retrieval, cross-modal hashing, semantic-graph, online learning, binary representations

1 INTRODUCTION

With the explosive growth of the Internet, increasing people would like to search information by both image and keyword. But due to the too much semantic information and complex relationship between objects, cross-modal scene retrieval is still a challenging problem. Lots of efforts have been paid to cross-modal scene retrieval in recent years [10, 15, 31–33, 37].

The binary representation or hashing method has been one of the promising techniques for large-scale cross-modal retrieval [2, 6, 17, 21, 43] for its low time and storage costs. Multi-modal data is common in the mobile Internet. For instance, *Instagram* and *Flickr*, the most popular photo sharing social apps, have more than 500 million photos with text tags or comments. Considering that the users always intend to search images by some natural language descriptions or mine text messages by visual data, cross-modal hashing becomes an important problem.

As can be seen from Figure 1, in an indoor scene image, there are three kinds of objects: a girl, a window and the chairs. If these objects can be identified from the image, direct link between image and text can be established. More information can be further extracted by detecting the object attribute and analyzing the spatial layout. Better understanding of the scene image is not only to study with regard to objects, but also the relationships between them. An ideal cross-modal retrieval method should be able to model the graph structure of semantic information and measure the similarity across different modalities precisely.

Besides, multi-modal data become available continuously as streams in the real-world Internet. Therefore, the cross-modal hashing should be performed in an online manner. Existing hashing methods usually use batch mode to retrain new hash functions, which are less efficient for streaming data. Considering that a query or search job is required to be responded very quickly, adaptive optimization is necessary and important.

In this paper, a novel framework for online cross-modal scene retrieval is proposed. The main contributions include:

- We adopt the cross-modal hashing method based on the quantization loss across different modal domains.
- A novel semantic-graph model is proposed to measure the semantic correlation and similarity between multi-modality instances.
- A two-step optimization method is developed to handle the online streaming data, which consists of an off-line step optimizes the quantizer with binary code learning, and a stochastic gradient descent based on-line step.

We conduct extensive experiments on four popular real-world multi-modal datasets to exploit the performance of our proposed method. Experimental results show that our approach is competitive with state-of-the-art method.

The rest of this paper is organized as follows. Related work is briefly discussed in Section 2. The proposed framework for online cross-modal scene retrieval is presented in Section 3. Then the experimental results are shown in Section 4. Finally, we draw the conclusion in Section 5.

*indicates the corresponding author(liannan@buaa.edu.cn).

MM'17, October 23–27, 2017, Mountain View, CA, USA.

DOI: 10.1145/3123266.3123311

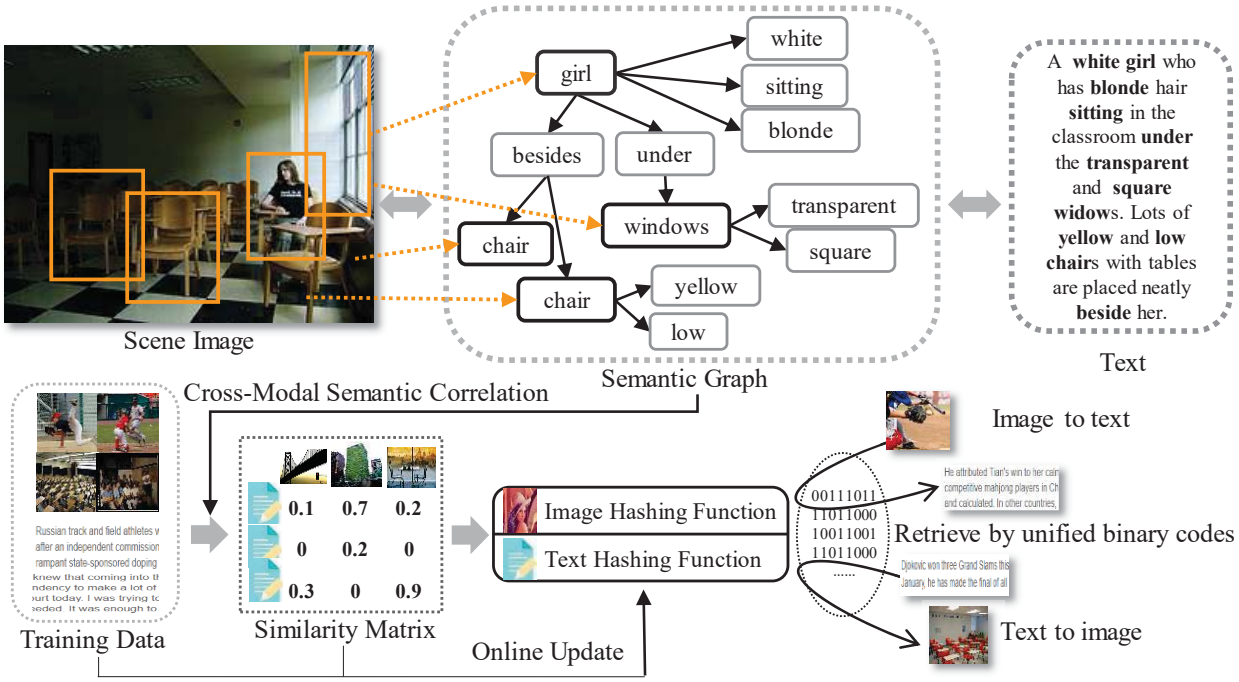

Figure 1: Overview of proposed framework. From left to right, we use training data and similarity matrix to learn hashing function for individual modal data, and the top process is measuring cross-modal semantic correlation with semantic graph.

2 RELATED WORK

Image Retrieval with Hashing. The locality-sensitive hashing (LSH) [9] is an important early work, in which binary codes are generated with random threshold by projecting data points to a random hyperplane. However, LSH requires a long hash code or a large number of hash table, which makes it inefficient. To address this issue, data-dependent hashing methods are proposed. Spectral hashing (SH) [35] uses Laplacian eigenfunction to compute the hash code based on uniform distribution. Iterative quantization (ITQ) [11] constructs hash codes by iteratively rotating the orthogonal projection of hash function. Kernelized locality-sensitive hashing (KLSH) [16] is an unsupervised learning method that can utilize unlabeled data. Recently, deep learning has been introduced to hashing based image retrieval [7, 40, 41].

Cross-Modal Hashing. The basic idea of cross-modal hashing is based on canonical correlation analysis (CCA) [11], which mapped two modal data into a common latent space by maximizing the correlation. Cross-modal similarity-sensitive hashing (CMSSH) [2] use Adaboost method to construct hash function for each modality. Cross-view hashing (CVH) [17] extends spectral hashing to cross-modal. Zhen et al. [42] do cross-modal hashing in a probabilistic way. Semantic correlation maximization (SCM) [39], which is inspired by Kernel-based supervised hashing, seamlessly integrate semantic labels into the learning procedure of hashing. Linear cross-modal hashing (LCMH) [44] preserves multi-modal similarity with different anchor graphs. Zhai et al. [38] propose a parametric method to learn individual projection matrix for each modal. Wang et al. [34] perform cross-model hashing by learning bridging mappings. Collective matrix factorization hashing (CMFH) [6] adopts latent collective matrix factorization to learn unified hash codes for different modalities. Latent semantic sparse hashing (LSSH) [43] adopts a sparse coding scheme to learn the latent semantic representation. Semantics-preserving hashing (SePH) [21] generates unified codes by minimizing the Kullback-Leibler divergence.

Online Hashing. The first work attempt to learn hash functions in an online manner is proposed by Huang et al. [12]. However, this work, which is named online kernel-based hashing (OKH), can only handle a pair of instances at one time. Online sketching hashing (OSH) [19] maintains a small size sketch of streaming data, which requires less computation and storage. Cakir and Sclaroff [3] propose an adaptive hashing approach for fast similarity search, which iteratively updated the hash functions based on stochastic gradient descent.

Graph-structured Representations. Fisher et al. [8] compare three dimensional scenes by graph kernels. Chang et al. [4] construct three dimensional scene by using a graph representation with language descriptions. Lin et al. [20] generate semantic graphs for video retrieval based on manually-defined rules. Johnson et al. [14] build a comprehensive scene graph using 5,000 images using manually labeled attributes.

3 PROPOSED APPROACH

As illustrated in Figure 1, our approach mainly consists of four parts: cross-modal binary representation, semantic graph across different modalities, the joint objective function and the online update method.

3.1 Cross-Modal Hashing

To achieve the cross-modal hashing across scene image and text label, we adopt the quantized correlation hashing method inspiring from [36]. Let $X \in R^{n\times d_x}$ and $Y \in R^{n\times d_y}$ be a coupled training set of two different modalities, where n is the sample number and d_x and d_y are the data dimensions of the two modalities respectively. The i^{th} row of X is denoted as x_i^T. Similarly the j^{th} row of Y is written as y_j^T. To simplify the formulation, we assume the data is zero-centered. The hash functions of two modalities are defined as $f(x_i) = sign(W_x^T x_i)$ and $g(y_j) = sign(W_y^T y_j)$ respectively [1], where $W_x \in R^{d_x\times c}$ and $W_y \in R^{d_y\times c}$ are two linear projections that can map two modalities into a linear subspace of equal dimension. What's more, we define S are the cross-modal similarity matrix (See Figure 1), which will be discussed later in Section 3.2. Based on the hash functions, we learn two kinds of binary codes $B_x \in \{-1,1\}^{n\times c}$ and $B_y \in \{-1,1\}^{n\times c}$ that of the same length c.

Cosine metric is used to measure the similarity between hash codes of different modalities,

$$\cos(f(x_i), g(y_j)) = \frac{f(x_i)^T g(y_j)}{\|f(x_i)\|_2 \|g(y_j)\|_2}. \tag{1}$$

Then we substitute the hash code with projection vector into Equation (1):

$$\cos(f(x_i), g(y_j)) \approx \frac{x_i^T W_x W_y^T y_j}{\sqrt{x_i^T W_x W_x^T x_i}\sqrt{y_j^T W_y W_y^T y_j}}.$$

Also, we adopt the subtraction operation instead of the ratio operation, which is inspired from the maximum margin criterion idea in [26]:

$$(x_i^T W_x W_y^T y_j - \sqrt{x_i^T W_x W_x^T x_i}\sqrt{y_j^T W_y W_y^T y_j}).$$

Taking the quantization loss [11] on each modality and the similarity constraint on cross-modality into consideration, the objective can be formulated as minimizing the following function:

$$\begin{aligned} &\min F(B_x, B_y, W_x, W_y) \\ &= (\|B_x - XW_x\|_F^2 + \|B_y - YW_y\|_F^2) \\ &- \alpha' \sum_{i,j} S_{ij}\Big(x_i^T W_x W_y^T y_j - \sqrt{x_i^T W_x W_x^T x_i}\sqrt{y_j^T W_y W_y^T y_j}\Big), \\ &s.t.\ W_x^T W_x = I,\ W_y^T w_y = I. \end{aligned} \tag{2}$$

Here, α' is a trade-off parameter balances the quantization loss and the cosine similarity. I is $c \times c$ identity matrix. Constraints $W_x^T W_x = I$ and $W_y^T w_y = I$ are used to make sure that W_x and W_y are orthogonal projections. Considering the inequality:

$$\frac{x_i^T W_x W_x^T x_i + y_j^T W_y W_y^T y_j}{2} \geq \sqrt{x_i^T W_x W_x^T x_i}\sqrt{y_j^T W_y W_y^T y_j},$$

[1] Here $sign(x) = \begin{cases} -1, & x \leq 0 \\ +1, & x > 0 \end{cases}$.

Equation (2) can be written in a matrix form:

$$\begin{aligned} &\min F(B_x, B_y, W_x, W_y) = (\|B_x - XW_x\|_F^2 + \|B_y - YW_y\|_F^2) \\ &- 2\alpha(tr(W_x^T X^T S Y W_y) - tr(W_x^T X^T L_x X W_x) \\ &- tr(W_y^T Y^T L_y Y W_y)), \end{aligned} \tag{3}$$

where $\alpha = \frac{1}{2}\alpha'$, and L_x and L_y are diagonal matrices represent the row-sum and column-sum of S respectively. Then we define

$$W = \begin{bmatrix} W_x \\ W_y \end{bmatrix},\ \widetilde{S} = \begin{bmatrix} \beta L_x & \alpha S \\ \alpha S^T & \beta L_y \end{bmatrix}, Z = \begin{bmatrix} X & \\ & Y \end{bmatrix},\ B = \begin{bmatrix} B_x, \\ B_y \end{bmatrix}.$$

The objective function can be written as

$$\begin{aligned} \min F(B, W) =& \|B - ZW\|_F^2 - tr(W^T Z^T \widetilde{S} Z W), \\ & s.t.\ W^T W = I. \end{aligned} \tag{4}$$

3.2 Semantic Graph

As illustrated in Figure 1, the proposed *Semantic Graph* is designed to be a bridge between scene image and corresponding text label, by which the similarity matrix S mentioned in Section 3.1 can be computed. It is constructed to describe object instances, attributes and inter-object relations that are regarded as important semantic features. Given a scene image these semantic elements can be extracted via object detection, while given a paragraph of text it can be archived by key word extraction. No matter what kind of input, the graph similarity is correlated to the similarity of scene category.

Graph Construction. Denote a set of object classes as $C = \{c_1, ..., c_{N(C)}\}$, attribute types set as $A = \{a_1, ..., a_{N(A)}\}$ and inter-object relationship set as $R = \{r_1, ..., r_{N(R)}\}$ respectively, an object in the scene image can be represented as $o_i = (c_i, a_i)$. Here $N(C)$, $N(A)$ and $N(R)$ are the numbers of object class, attribute type and inter-object relationship respectively. We can define a semantic graph as $G = (O, E)$, where $O = \{o_1, ..., o_i\}$ is a set of objects and $E = \{O \times R \times O\}$ is a set of edges between different objects. In this work, the semantic graph of the scene is constructed by adopting the *real-world scene graphs* [14], which consists of over 5,000 images, 93,000 objects, 110,000 types of attributes, 112,000 types of relationships.

Mapping from Text to Graph. Since the keywords can represent rich semantic information of a scene, we extract noun, verb and adjective and preposition from text by the Stanford parser [22]. Let $T = \{k_1, ..., k_{N(K)}\}$ be a paragraph of text, where k_i denotes the i^{th} keyword and $N(k)$ is the number of keywords. We denote the mapping from text T to semantic graph $G = (O, E)$ as $\{\psi : k \to o, e\}$. Because the graph is represented by natural language, the output of mapping is set to binary, i.e. if the keywords is matched with the part of semantic-graph $\{\psi : k \to o, e\} = 1$, otherwise $\{\psi : k \to o, e\} = 0$. The matching possibility between text T and graph G is:

$$\begin{aligned} P(\psi|T, G) = \frac{1}{N(C) + N(A) + N(R)}\Big(&\sum_{i=1}^{N(k)}\sum_{j=1}^{N(C)} 1\{\psi : k_i \to c_j\} \\ + \sum_{i=1}^{N(k)}\sum_{j=1}^{N(A)} 1\{\psi : k_i \to a_j\} &+ \sum_{i=1}^{N(k)}\sum_{j=1}^{N(R)} 1\{\psi : k_i \to r_j\}\Big). \end{aligned} \tag{5}$$

Mapping from Image to Graph. The mapping from scene image to the semantic graph is achieved by detecting the objects. Then faster regional convolutional neural networks (faster R-CNN) [24, 30] is used to obtain the class, attribute and location of appeared objects. The results are further converted to possibility scores by applying the random forest [1]. Let κ be the *grounding parts* that represents the part of graph correspondence to the scene image. We model the distribution of all possible grounding parts in new image by conditional random field (CRF) [18]. The grounding parts of the maximum possibility is obtained by maximum a posteriori inference.

For an input scene image, we can get a set of bounding boxes $B = \{b_1, ..., b_n\}$. We draw a mapping $\{\phi : \kappa_i \rightarrow B\}$ from the grounding parts of the semantic graph to bounding boxes. So the distribution of possible grounding parts is:

$$P(\phi \mid G, B) = \prod_{o \in O} P(\phi_o \mid o) \prod_{(o,r,o') \in E} P(\phi_o, \phi_{o'} \mid o, r, o'). \quad (6)$$

Depending on $P(\phi_o \mid o) = P(o \mid \phi_o)P(\phi_o)/P(o)$, we have the simplified objective

$$\phi^* = \arg\max_{\phi} \prod_{o \in O} P(o \mid \phi_o) \prod_{(o,r,o') \in E} P(\phi_o, \phi_{o'} \mid o, r, o'). \quad (7)$$

The term $P(o \mid \phi_o)$ is regarded as the unary energy. Draw $o = (c, A)$ into it we have

$$P(o \mid \phi_o) = P(c \mid \phi_o) \prod_{a \in A} P(a \mid \phi_o), \quad (8)$$

where $P(\phi_o, \phi_{o'} \mid o, r, o')$ represents the binary energy that models the bounding box pair in the scene image. Let the bounding box coordinates be $\kappa_o = (x, y, w, h)$ and $\kappa_{o'} = (x', y', w', h')$ respectively, then the spatial relationship between them can be defined as

$$R(\phi_o, \phi_{o'}) = \big((x - x')/w, (y - y')/h, w'/w, h'/h\big).$$

For a specific object class c, Gaussian mixture model (GMM) [45] is used to get $P(R(\phi_o, \phi_{o'}) \mid c, r, c')$. After that, random forest is used to transform the GMM density value at $R(\phi_o, \phi'_o)$ to a probability $P(\phi_o, \phi_{o'} \mid o, r, o')$.

Semantic Correlation. As previously stated, we achieve the final cross-modal semantic correlation S_{ij} between the i^{th} image and the j^{th} text instance by

$$S_{ij} = P(\psi \mid G, T)P(\phi \mid G, B). \quad (9)$$

We further conduct a normalization step, i.e., $S_{ij} = \frac{s_{ij}}{\sum s_{ij}}$ to force its value in the range of $[0, 1]$.

3.3 Joint Objective Function

After constructing the semantic graph across scene image and text data, we get the original similarity matrix S. However, most conventional hashing methods quantified the similarity between two instance points to $(+1, -1)$ or other arbitrary values, which could not describe the correct relationship among training data. To address this problem, we learn S automatically through a joint framework integrating the cross-modal hashing and semantic graph.

Based on Equation (4), the joint objective can be formulated as

$$\begin{aligned} &\min F(B, W, S) := \ell_1(F, S) + \mu\{\ell_2(X, S) + \ell_3(Y, S)\} \\ &+ \xi r(S) + \nu(\|B - ZW\|_F^2 - tr(W^T Z^T \widetilde{S} Z W)), \\ &\text{s.t.} \quad W^T W = I, S \geq 0. \end{aligned} \quad (10)$$

Here $\ell_1(F, S)$, $\ell_2(X, S)$ and $\ell_3(Y, S)$ are the loss functions to measure the smoothness of S on the scene semantic features F, image feature X and text feature Y, respectively. $r(S)$ is regularization terms w.r.t. S. μ, ξ and ν are the balancing parameters. We define

$$\begin{aligned} &\ell_1(F, S) = \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij},\ \ell_2(X, S) = \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij}, \\ &\ell_3(Y, S) = \sum_{i,j} \|y_i - y_j\|_2^2 s_{ij},\ r(S) = \|S\|_F^2, \end{aligned} \quad (11)$$

where x and y represent raw features of scene image and text respectively, and f_i and f_j are the semantic features of data points x_i and x_j. The semantic feature consists of the object category, attributes and relationships in scene semantic graph, which is a 5,000 dimensional vector [14].

3.4 Online Optimization

We develop a two-step optimization strategy. In the first step S, W and B are iteratively updated off-line, while in the second step W is respectively optimized for each modal online.

3.4.1 Off-line Step. **Optimizing S.** By fixing W and B, we could update S according to Equation (10). The simplified objective function as

$$\begin{aligned} &\min_S \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij} + \mu \sum_{i,j} (\|x_i - x_j\|_2^2 - \|y_i - y_j\|_2^2) s_{ij} \\ &+ \xi \|S\|_F^2 + \nu(\|B_x - XW_x\|_F^2 + \|B_y - YW_y\|_F^2 - \alpha' \sum_{i,j} S \\ &(x_i^T W_x W_y^T y_i - \sqrt{x_i^T W_x W_x^T x_i} \sqrt{y_j^T W_y W_y^T y_j})) \\ &\Rightarrow \min_S \sum_i tr\big((a_i + \mu(b_i + c_i))s_i^T + \xi s_i s_i^T - \nu\alpha' s_i^T d_i\big) + \nu C. \end{aligned} \quad (12)$$

Here $a_i = \{a_{ij}, 1 \leq j \leq N\}$ and $a_{ij} = \|f_i - f_j\|_2^2$,$b_i = \{b_{ij}, 1 \leq j \leq N\}$ and $b_{ij} = \|x_i - x_j\|_2^2$; $c_{ij} = \|y_i - y_j\|_2^2$; $d_i = \{d_{ij}, 1 \leq j \leq N\}$ and $d_{ij} = (x_i^T W_x W_y^T y_i - \sqrt{x_i^T W_x W_x^T x_i} \sqrt{y_j^T W_y W_y^T y_j})$. Since $C = \|B_x - XW_x\|_F^2 + \|B_y - YW_y\|_F^2$ is a constant, Equation (12)) can be reformulated as

$$\begin{aligned} &\min_S \sum_i tr\Big((\xi s_i s_i^T - (\nu\alpha' d_i - (a_i + \mu b_i + \mu c_i))s_i^T)\Big) \\ &\Rightarrow \min_S \sum_i \|s_i - \frac{\nu\alpha' d_i - (a_i + \mu b_i + \mu c_i)}{\xi}\|_2^2, \end{aligned} \quad (13)$$

which can be solved by accelerated projected gradient method [28].

Optimizing B. By fixing W and S, Equation (4) turns out to be a regularized least-square problem:

$$\min F(B) := \|B\|_F^2 + \|Z\|_F^2 - 2tr(BW^T Z^T). \quad (14)$$

Because W and Z are fixed, the hash code B_x and B_y have the same sign as ZW, this problem has a close-form solution:

$$B = sign(ZW). \quad (15)$$

Algorithm 1 Online Optimization

Input:

Two-modality streaming data X and Y, individual modal pairs $\{(x_i^t, x_j^t)\}_{t=1}^T$ and $\{(y_i^t, y_j^t)\}_{t=1}^T$, original similarity matrix S of semantic-graph between two modalities, code length c, initialize W_o, iteration number N, trade-off parameters: μ, ξ, ν, α, β, λ, ρ, η^t;

Output:

Projection matrix W, similarity matrix S_{ij} of semantic-graph and hash codes B;

Off-line Step

for $n = 1 \rightarrow N$ **do**
  Optimize S when fixing W and B using Eq.(13) ;
  Optimize B when fixing W and S using Eq.(15) ;
  Iteratively optimize W when fixing B and S using Eq.(17);
end for

On-line Step

for $t = 1 \rightarrow T$ **do**
  Compute binary codes $f(x_i^t)$, $f(x_j^t)$;
  Compute loss $l_h(f(x_i), f(x_j); W_x)$ according to Eq.(18);
  if $\lceil l_h \rceil \neq 0$ **then**
    Compute $\nabla_W l(x_i, x_j, W^t)$;
    compute and sort Λ;
    $j \leftarrow$ the incorrect first $\lceil l_h \rceil$ by sorted Λ;
    $\nabla_W l(:, j) \leftarrow 0$;
    $W^{t+1} \leftarrow W^t$ - $\eta^t \nabla_W l(x_i, x_j, W^t)$;
  else
    $W^{t+1} \leftarrow W^t$
  end if
end for
Repeat step 6-18 for y.

Optimizing W. To solve with fixed B and S, we introduced Lagrangian multipliers with orthogonal constraint, Equation (4) becomes

$$\begin{aligned} L(W, \lambda) &= F(W) - \frac{1}{2} tr(\lambda(W^T W - I)) \\ &= \|B - ZW\|_F^2 - tr(W^T Z^T \widetilde{S} Z W) \\ &- \frac{1}{2} tr(\lambda(W^T W - I)) \\ &= \|B\|_F^2 + tr(W^T Z^T Z W) - 2tr(B W^T Z^T) \\ &- tr(W^T Z^T \widetilde{S} Z W) - \frac{1}{2} tr(\lambda(W^T W - I)). \end{aligned} \tag{16}$$

To make sure that the gradient of Equation (16) is zero, i.e. $\frac{\partial L(W,\lambda)}{\partial W} = \frac{\partial F(W)}{\partial W} - W\lambda = 0$, we define $G = \frac{\partial F(W)}{\partial W}$. Consequently we have $\lambda = W^T G = G^T W$. Let $A = GW^T - G^T W$, W can be updated by Crank-Nicolson-like scheme [27],

$$W^{t+1} = QW^{(t)}, \quad Q = (I + \frac{\tau}{2} A)^{-1}(I - \frac{\tau}{2} A). \tag{17}$$

Here τ is the step size.

3.4.2 On-line Step. To tackle streaming data, inspired by [3] we adopt an online learning algorithm, which can update the hash function fast with stochastic gradient descent. When a pair of new instances are inputed, we could update W_x and W_y in the hash function respectively.

Taking modal data X for instance, if the label of X and the similarity matrix $s_{ij}^x \in \{-1, 1\}$ of its elements are known. We employ the squared error loss and take an orthogonality regularizer into account,

$$l(f(x_i), f(x_j); W_x) = (f(x_i)^T f(x_j) - B_x s_{ij}^x)^2 + \frac{\lambda'}{4} \|W^T W - I\|_F^2.$$

W can be updated by online SGD,

$$W_{t+1} \leftarrow W^t - \eta^t \nabla l(f(x_i), f(x_j); W_x).$$

In order to determine which hash function need to be updated, we also develop a hinge-like loss function of [12].

$$l_h(f(x_i), f(x_j)) = \begin{Bmatrix} \max(0, d_H^x - (1-\rho)B_x) & s_{ij} = 1 \\ \max(0, \rho B_x - d_H^x) & s_{ij} = -1 \end{Bmatrix}, \tag{18}$$

where d_H^x is the hamming distance and $\rho \in [0, 1]$ is used to balance the loss. Moreover, we develop $\Omega = \{max(\frac{|f_1(x_i)|}{\|\overline{w_1}\|}, \frac{|f_1(x_j)|}{\|\overline{w_1}\|}), \ldots, max(\frac{|f_{B_x}(x_i)|}{\|\overline{w_{B_x}}\|}, \frac{|f_{B_x}(x_j)|}{\|\overline{w_{B_x}}\|})\}$, where $\overline{w} = [w, w_0]$. Then we can sort the set Ω in descending order and the first $\lceil l_h \rceil$ means the incorrect hash function.

W_y can be updated similarly by the above-described method. Whole procedure of the two-step optimization is summarized in Algorithm 1.

4 EXPERIMENTS

In this section, we conducted experiments to demonstrate the effectiveness of our framework on four large-scale benchmarks: Wiki [23], NUS-WIDE [5], MIRFlickr [13], LM+Sun [29]. Two series of experiments are performed, i.e. image-to-text scene retrieval and text-to-image scene retrieval respectively.

4.1 Dataset and Experiment Setup

Datasets Wiki dataset [23] consists of 2,866 text documents that crawled from the Wikipedia's featured articles with manual marked image-text pairs. The pair is further divided into ten categories. Each document is represented by a 10-dimensional feature vector through latent Dirichlet allocation (LDA), and each image is described by a 128-dimensional SIFT feature vector. Following the experiment settings in [23], we choose 2173 image-text pairs (about 75%) and 693 image-document pairs (about 25%) left as training set and queries respectively.

NUS-WIDE [5] contains 269,648 images crawled from Flickr with associated text tags. The image-text pairs are divided into 81 categories. Ten categories with largest sample number, which consist of 186,577 image-text pairs, are used in the experiments. In the experiments, an image is represented by a 500-dimensional bag-of-visual-words feature, and corresponding text is described by a 1,000 dimensional vector. Following the settings in [5], we further sample 99% of image-text pairs as the training set, while remaining pairs are used as queries.

MIRFlickr [13] consists 25,000 images with corresponding text tags collected from Flickr. Each image has several associated text labels from 38 unique categories. As in [25], every image is described by a 3857-dimensional feature and corresponding text is

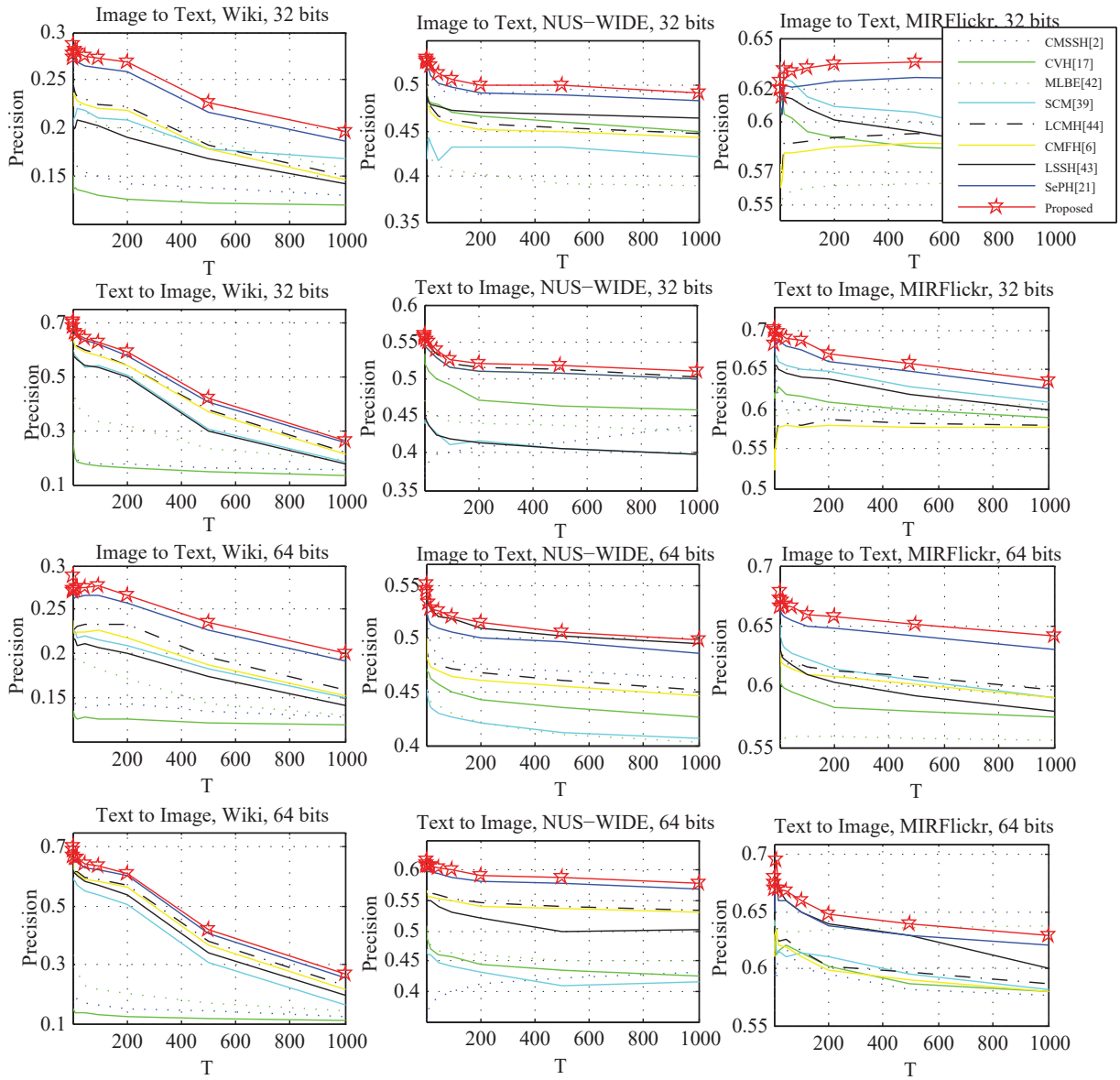

Figure 2: Curves of the top T precision on the Wiki, NUS-WIDE and MIRFlickr varying the code length.

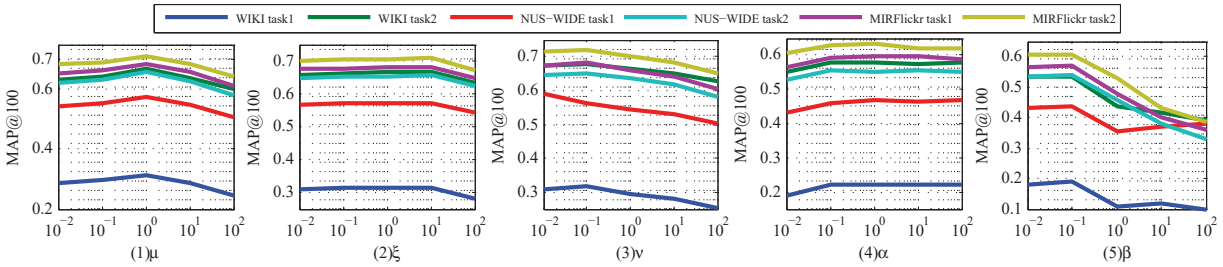

Figure 3: Parameter analysis of our approach with respect to μ, ξ, ν, α and β over image-to-text (task1) and text-to-image (task2) on Wiki, NUS-WIDE and MIRFlickr with code length of 128 bits.

Table 1: Comparisons of mean average precision(mAP) on Wiki, NUS-WIDE and MIRFLickr with code length 16, 32, 64 and 128 bits. ∗ denotes the results are cited from corresponding reference. † means the results are obtained by using codes provided by the authors. Results based on our implementation is marked by ‡.

Task	Method	Wiki				NUS-WIDE				MIRFlickr			
		16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits	16 bits	32 bits	64 bits	128 bits
Img to Txt	CMSSH[2]†	0.1853	0.1732	0.1653	0.1535	0.4028	0.3986	0.3886	0.3801	0.5734	0.5741	0.5716	0.5702
	CVH[17]‡	0.1947	0.1798	0.1732	0.1676	0.3652	0.4154	0.4611	0.4496	0.6075	0.6177	0.6156	0.6078
	MLBE[42]∗	0.2561	-	-	-	-	-	-	-	0.6608	-	-	-
	MLBE[42]†	0.2537	0.2747	0.2599	0.2489	0.4472	0.4540	0.4703	0.4699	0.5689	0.5766	0.5799	0.5683
	SCM[39]∗	0.2393	0.2419	-	-	0.4385	0.4390	-	-	-	-	-	-
	SCM[39]†	0.2393	0.2379	0.2419	0.2605	0.4385	0.4397	0.4390	0.4407	0.6236	0.6325	0.6467	0.6494
	LCMH[44]∗	0.273	-	-	-	0.562	-	-	-	-	-	-	-
	LCMH[44]‡	0.2531	0.2572	0.2699	0.2706	0.4657	0.4729	0.4789	0.4798	0.6382	0.6399	0.6407	0.6578
	CMFH[6]∗	0.2538	0.2582	0.2619	0.2648	**0.5591**	**0.5698**	**0.5780**	**0.5837**	0.6480	0.6597	0.6693	0.6752
	CMFH[6]†	0.2538	0.2582	0.2619	0.2648	0.4391	0.4298	0.4280	0.4237	0.6480	0.6597	0.6693	0.6752
	LSSH[43]∗	0.2330	0.2340	0.2387	0.2340	0.4933	0.5006	0.5069	0.5084	-	-	-	-
	LSSH[43]†	-	-	-	-	-	-	-	-	0.5776	0.5781	0.5807	0.5672
	SePH[21]∗	0.2787	0.2956	0.3064	0.3134	0.5421	0.5499	0.5537	0.5601	0.6723	0.6771	0.6783	0.6817
	SePH[21]†	0.2791	0.2962	0.3053	0.3131	0.5412	0.5481	0.5526	0.5608	0.6729	0.6773	0.6785	0.6813
	Our Method	**0.2802**	**0.3078**	**0.3196**	**0.3291**	0.5585	0.5593	0.5617	0.5712	**0.6756**	**0.6875**	**0.6897**	**0.6906**
Txt to Img	CMSSH[2]†	0.1621	0.1592	0.1559	0.1541	0.3898	0.3776	0.3752	0.3676	0.5693	0.5738	0.5711	0.5681
	CVH[17]‡	0.1196	0.1043	0.1019	0.1001	0.3627	0.4030	0.4341	0.4259	0.6038	0.6036	0.6011	0.5968
	MLBE[42]∗	0.3209	-	-	-	-	-	-	-	0.5970	-	-	-
	MLBE[42]†	0.3336	0.3693	0.3597	0.3568	0.4652	0.4889	0.5050	0.5087	0.6135	0.6277	0.6359	0.6397
	SCM[39]∗	0.2325	0.2452	-	-	0.5147	0.5105	-	-	-	-	-	-
	SCM[39]†	0.2325	0.2454	0.2452	0.2574	0.4273	0.4265	0.4259	0.4362	0.6142	0.6207	0.6302	0.6319
	LCMH[44]∗	0.423	-	-	-	0.621	-	-	-	-	-	-	-
	LCMH[44]‡	0.3590	0.3551	0.3553	0.3690	0.4729	0.5128	0.5159	0.5189	0.6234	0.6389	0.6456	0.6499
	CMFH[6]∗	0.6116	0.6298	0.6398	0.6477	**0.6614**	**0.6921**	**0.7164**	**0.7185**	0.6174	0.6241	0.6311	0.6340
	CMFH[6]†	0.6116	0.6298	0.6398	0.6477	0.4614	0.4561	0.4524	0.4465	0.6174	0.6241	0.6311	0.6340
	LSSH[43]∗	0.5571	0.5743	0.5710	0.5577	0.6250	0.6578	0.6823	0.6913	-	-	-	-
	LSSH[43]†	-	-	-	-	-	-	-	-	0.5911	0.5921	0.5939	0.5955
	SePH[21]∗	**0.6318**	0.6577	0.6646	**0.6709**	0.6302	0.6425	0.6506	0.6580	0.7197	0.7271	0.7309	0.7354
	SePH[21]†	0.6216	0.6571	0.6639	0.6702	0.6311	0.6419	0.6508	0.6577	0.7199	0.7268	0.7302	0.7349
	Our Method	0.6316	**0.6627**	**0.6773**	0.6691	0.6429	0.6516	0.6638	0.6698	**0.7228**	**0.7356**	**0.7378**	**0.7392**

Table 2: Comparisons on LM+Sun dataset.

Task	Method	mAP@100				Training time(seconds)
		8 bits	16 bits	32 bits	64 bits	32 bits
Img to Txt	OKH[12]+MLBE[42]	0.2934	0.3598	0.4658	0.5822	170
	OSH[19]+MLBE[42]	0.2464	0.2978	0.4277	0.5289	312
	Our Method	**0.3055**	**0.3758**	**0.4699**	**0.5962**	**150**
Txt to Img	OKH[12]+MLBE[42]	0.3867	0.4489	0.5768	0.6523	130
	OSH[19]+MLBE[42]	0.3120	0.3690	0.5188	0.5709	253
	Our Method	**0.3928**	**0.4769**	**0.5788**	**0.6721**	**112**

represented by a 2,000 dimensional vector that indicates its occurrence. We randomly take 5% image-text pairs as the query set, and the rest pairs are used as the training set.

The training and test set of LM+Sun [29] includes 45,676 and 500 scene images respectively, including both indoor and outdoor scenes. There are 232 kind of scenes in total. Following the experiment setups in [29], 90 percents of all image-tag pairs as training pairs, and the remaining are queries.

Evaluation Metrics We use mean average precision (mAP) as the performance measure. Given a query and a set of T retrieved items, the average precision (AP) is defined as

$$AP = \frac{1}{L}\sum_{t=1}^{T} P(t)\delta(t),$$

where L is the number of true neighbours in the retrieved set, $P(t)$ denotes the precision of top t retrieved items, and $\delta(t) = 1$ if the t-th

retrieved item is a true neighbour and $\delta(t) = 0$ otherwise. Ground truth neighbors are defined as those pairs which share at least one label. Given a query set of size Q, the mAP is defined as the mean of average precision scores for all the queries in query set

$$mAP = \frac{1}{Q}\sum_{i=1}^{Q} AP(q_i).$$

In the experiments, mAP is obtained by repeating the retrieval ten times and averaging the results.

Compared Methods First, we compare the proposed approach with eight cross-modal image retrieval methods: Cross-Modal Similarity-Sensitive Hashing (CMSSH) [2], Cross-View Hashing (CVH) [17], Multi-modal Latent Binary Embedding (MLBE) [42], Semantic Correlation Maximization (SCM) [39], Linear Cross-Modal Hashing (LCMH) [44], Collective Matrix Factorization Hashing (CMFH) [6], Latent Semantic Sparse Hashing (LSSH) [43], and Semantics-Perserving Hashing (SePH) [21]. To demonstrate the benefit in hashing, two online hashing methods combine with MLBE, i.e. Online Kernel-based Hashing (OKH) [12] and Online Sketching Hashing (OSH) [19], are also compared[2].

4.2 Results and Analysis

Wiki, NUS-WIDE and MIRFlickr Although the proposed method is designed for scene retrieval, it can be also applied to general text-to-image/image-to-text retrieval tasks by simply treating input image/text as a kind of scene. Considering that scene dataset with multiple modalities is relatively rare, comparisons with general cross-modal image retrieval method can be a good performance reference. Therefore, we first conduct experiments on Wiki, NUS-WIDE and MIRFlickr. The results and comparisons can be found in Figure 2 and Table 1.

Besides the results based on our implementation, those reported in corresponding references are also shown in Table 1. It should be pointed out that there are some differences of experiment setting. In [42] and [39], the authors use 80% of the data for training and the remaining 20% to form the query set on the Wiki dataset. In [42] on Flickr and [39] on NUS-WIDE, 99% of the data is used as gallery set and the rest 1% forms the query set. In [6], 75% and 25% of the data are used as dataset and query respectively on Wiki and MIRFlickr. For NUS-WIDE, the dataset division is similar to [39]. In [43] and [21], the dataset division is 75% vs. 25% on Wiki and 99% vs. 1% on NUS-WIDE.

As can be seen, the proposed method achieves the best results in the image-to-text retrieval on the Wiki dataset. In text-to-image retrieval its performance is comparable to SePH method [21]. On the NUS-WIDE dataset, our method get the second place in both image-to-text and text-to-image retrievals. As reported in [6], CMFH method gets the best results. However, our method is at the first place if based on our implementation. In MIRFlickr dataset our approach consistently outperforms the compared methods. Generally speaking, we can reach a conclusion that the proposed method is competitive among the state-of-the-art as a general image retrieval method.

[2]In the experiments, the parameter settings of above-mentioned methods are adopted from corresponding papers.

It should be pointed out that the proposed method is only designed for scene retrieval scenario. The results imply that better scene understanding can effectively enhance the performance of general cross-modal image retrieval. Although not designed for scene retrieval, Wiki, NUS-WIDE and MIRFlickr include quite a number of scene image/text pairs. The experimental results shown in Figure 2 and Table 1 can be a good reference of cross-modal scene retrieval.

LM+Sun Besides general image/text retrieval, we also conduct experiments on LM+Sun, which only consists of scene data. Results in term of mean average precision of first 100 retrieved items (mAP@100) are shown in Table 2. Obviously, our method consistently outperforms both OKH+MLBE and OSH+MLBE. When the code length decreases, the performance drops faster than it in Wiki, NUS-WIDE and MIRFlickr. This phenomenon shows that scene is a complex concept that requires more details for effective representation. For an online algorithm, update time is another important performance indicator. We compare the training time with 32 bits code in Table 2. The proposed method achieves lowest time cost, which implies that the proposed method is both effective and efficient.

Parameter Sensitivity Analysis. Parameters have a great influence on the performance. To investigate their sensitivity, experiments of different parameter settings are also conducted. As described in Section 3.4, in the off-line step, five parameters affect the performance of our framework, i.e. μ, ξ, ν, α and β respectively. α controls the trade-off between hash function learning and quantization and β is a regularizer coefficient. The default setting for the parameters are: $\mu = 1$, $\xi = 10$, $\nu = 10^{-1}$, $\alpha = 0.01$, $\beta = 0.01$. We tune these five parameters in the range of $\{0.01, 0.1, 1, 10, 100\}$. The results are shown in Figure 3. We find that when $\mu = 1$, $\xi = 10$ and $\nu = 0.1$, the performance reach its peak, which means balancing the semantic-graph and feature achieve better results. Also, we can see when $\alpha > 0.1$ and $0.01 < \beta < 0.1$ on Wiki, NUS-WIDE and MIRFlickr datasets, the results of hashing methods seem not sensitive to parameter. In the on-line step, parameters α' is designed to balance the hash function loss, which is fixed to 0.6.

5 CONCLUSIONS

In this paper, we introduce a novel framework for online cross-modal scene retrieval with binary representations and semantic-graph. We adopt the cross-modal hashing based on the quantized correlation, and measure the semantic agreement and similarity of semantic-graph for each instance. The problem is effectively optimized by a two-step strategy. Extensive experiments on four datasets indicate that our approach outperforms the state-of-the-art. It would be a promising future work to develop graph kernel with deep learning to retrieve scene images.

6 ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 61573045 and by the Foundation for Innovative Research Groups through the National Natural Science Foundation of China and Grant 61421003.

REFERENCES

[1] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Image Classification using Random Forests and Ferns. In *ICCV*. 1–8.

[2] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, Vol. 1. 5.

[3] Fatih Cakir and Stan Sclaroff. 2015. Adaptive hashing for fast similarity search. In *ICCV*. 1044–1052.

[4] Angel X Chang, Manolis Savva, and Christopher D Manning. 2014. Learning Spatial Knowledge for Text to 3D Scene Generation. In *EMNLP*. 2028–2038.

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *ACM international conference on image and video retrieval*. ACM, 48.

[6] Guiguang Ding, Yuchen Guo, and Jile Zhou. 2014. Collective matrix factorization hashing for multimodal data. In *CVPR*. 2075–2082.

[7] Venice Erin Liong, Jiwen Lu, Gang Wang, Pierre Moulin, and Jie Zhou. 2015. Deep hashing for compact binary codes learning. In *CVPR*. 2475–2483.

[8] Matthew Fisher, Manolis Savva, and Pat Hanrahan. 2011. Characterizing structural relationships in scenes using graph kernels. In *SIGGRAPH*, Vol. 30. 34.

[9] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity search in high dimensions via hashing. In *VLDB*, Vol. 99. 518–529.

[10] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. *IJCV* 106, 2 (2014), 210–233.

[11] Yunchao Gong and Svetlana Lazebnik. 2011. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*. 817–824.

[12] Long-Kai Huang, Qiang Yang, and Wei-Shi Zheng. 2013. Online Hashing. In *IJCAI*.

[13] Mark J Huiskes and Michael S Lew. 2008. The MIR flickr retrieval evaluation. In *MIR*. 39–43.

[14] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*. 3668–3678.

[15] C Kang, S Xiang, S Liao, and C Xu. 2015. Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval. *TMM* 17, 3 (2015), 370–381.

[16] Brian Kulis and Kristen Grauman. 2009. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*. 2130–2137.

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

[18] John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data. In *ICML*, Vol. 3. 282–289.

[19] Cong Leng, Jiaxiang Wu, Jian Cheng, Xiao Bai, and Hanqing Lu. 2015. Online sketching hashing. In *CVPR*. 2503–2511.

[20] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. 2014. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*. 2657–2664.

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

[22] DKCD Manning. 2003. Natural language parsing. In *NIPS*, Vol. 15. 3.

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

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[25] Ruslan Salakhutdinov and Geoffrey E Hinton. 2009. Deep Boltzmann Machines. In *AISTATS*, Vol. 1. 3.

[26] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *CVPR*. 2160–2167.

[27] Gordon D Smith. 1965. Numberical Solution of Partial Different Equation.

[28] Jingkuan Song, Lianli Gao, Feiping Nie, Heng Tao Shen, Yan Yan, and Nicu Sebe. 2016. Optimized Graph Learning Using Partial Tags and Multiple Features for Image and Video Annotation. *TIP* 25, 11 (2016), 4999–5011.

[29] Joseph Tighe and Svetlana Lazebnik. 2010. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*. 352–365.

[30] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *IJCV* 104, 2 (2013), 154–171.

[31] Adrian Ulges and Christian Schulze. 2011. Scene-based image retrieval by transitive matching. In *ACM International Conference on Multimedia Retrieval*. 47.

[32] Han Wang, Wei Liang, Xinxiao Wu, and Peng Teng. 2013. Scene image retrieval via re-ranking semantic and packed dense interestpoints. *Neurocomputing* 119, 16 (2013), 65–73.

[33] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan. 2013. Learning Coupled Feature Spaces for Cross-Modal Matching. In *ICCV*. 2088–2095.

[34] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. 2015. Lbmch: Learning bridging mapping for cross-modal hashing. In *SIGIR*. 999–1002.

[35] Yair Weiss, Antonio Torralba, and Rob Fergus. 2009. Spectral hashing. In *NIPS*. MIT Press, 1753–1760.

[36] Botong Wu, Qiang Yang, Wei-Shi Zheng, Yizhou Wang, and Jingdong Wang. 2015. Quantized correlation hashing for fast cross-modal search. In *IJCAI*. 25–31.

[37] Fei Wu, Xinyan Lu, Zhongfei Zhang, Shuicheng Yan, Yong Rui, and Yueting Zhuang. 2013. Cross-media semantic representation via bi-directional learning to rank. In *ACM Multimedia*.

[38] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, and Wen Gao. 2013. Parametric Local Multimodal Hashing for Cross-View Similarity Search. In *IJCAI*.

[39] Dongqing Zhang and Wu-Jun Li. 2014. Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization. In *AAAI*, Vol. 1. 7.

[40] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama. 2016. Efficient Training of Very Deep Neural Networks for Supervised Hashing. In *CVPR*. 1487–1495.

[41] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *CVPR*. 1556–1564.

[42] Yi Zhen and Dit-Yan Yeung. 2012. A probabilistic model for multimodal hash function learning. In *SIGKDD*. 940–948.

[43] Jile Zhou, Guiguang Ding, and Yuchen Guo. 2014. Latent semantic sparse hashing for cross-modal similarity search. In *SIGIR*. 415–424.

[44] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. 2013. Linear cross-modal hashing for efficient multimedia search. In *ACM MM*. 143–152.

[45] Z. Zivkovic and F. Van Der Heijden. 2004. Recursive unsupervised learning of finite mixture models. *PAMI* 26, 5 (2004), 651.