Quantized Correlation Hashing for Fast Cross-Modal Search

Botong Wu[†‡◇], Qiang Yang[†], Wei-Shi Zheng[†§*], Yizhou Wang[‡], Jingdong Wang[♯]
[†] School of Information Science and Technology, Sun Yat-sen University, China
[‡] Nat'l Eng. Lab. for Video Technology, Cooperative Medianet Innovation Center, Sch'l of EECS, Peking University
[◇] Collaborative Innovation Center of High Performance Computing, National University of Defense Technology
[§] Guangdong Provincial Key Laboratory of Computational Science [♯] Microsoft Research Asia, China
{issbotongwu,mmmyqmmm}@gmail.com,wszheng@ieee.org,yizhou.wang@pku.edu.cn,jingdw@microsoft.com

Abstract

Cross-modal hashing is designed to facilitate fast search across domains. In this work, we present a cross-modal hashing approach, called quantized correlation hashing (QCH), which takes into consideration the quantization loss over domains and the relation between domains. Unlike previous approaches that separate the optimization of the quantizer independent of maximization of domain correlation, our approach simultaneously optimizes both processes. The underlying relation between the domains that describes the same objects is established via maximizing the correlation between the hash codes across the domains. The resulting multi-modal objective function is transformed to a unimodal formalization, which is optimized through an alternative procedure. Experimental results on three real world datasets demonstrate that our approach outperforms the state-of-the-art multi-modal hashing methods.

1 Introduction

Multi-modal data becomes more and more popular in our daily life. For instance, on webpages, objects are often described with images or videos surrounded by text. With the interests on searching multi-modal data (e.g. using images to query texts), cross-modal retrieval becomes an emergent issue. A lot of works have been done in this field [Li *et al.*, 2003a; Rasiwasia *et al.*, 2010; Gong *et al.*, 2014; Hwang and Grauman, 2012; Sharma *et al.*, 2012]. Recent efforts focus on studying efficient search methodologies over large databases. Inspired by the fast search based on hashing algorithms [Datar *et al.*, 2004; Weiss *et al.*, 2009; Wang *et al.*, 2010; Liu *et al.*, 2012; Norouzi and Blei, 2011], which have witnessed great success in single-modal search, cross-modal hashing[Masci *et al.*, 2013; Bronstein *et al.*, 2010; Zhang and Li, 2014; Zhou *et al.*, 2014; Zhen and Yeung, 2012a] or cross-view hashing[Kim and Choi, 2013; Quadrianto and Lampert, 2011; Song *et al.*, 2013; Zhai *et al.*, 2013; Zhang *et al.*, 2011; Zhou *et al.*, 2014] have been attracting a lot.

*Corresponding Author

Most cross-modal hashing methods are similarity based (for example, Cross Modality Similarity Sensitive Hashing(CMSSH) [Bronstein *et al.*, 2010] and Semantic Correlation Maximization(SCM) [Zhang and Li, 2014]) or distance-based (e.g. Cross View Hashing(CVH) [Kumar and Udupa, 2011] and Co-Regularized Hashing(CRH) [Zhen and Yeung, 2012a]). The objective of these methods is to purse "meaningful" hash codes of data points across domains, i.e., the distance between two codes - in the mapped hashing spaces - should be close if the respective data points are from the same class, and it should be far otherwise. Parametric Local Multimodal Hashing(PLMH) [Zhai *et al.*, 2013] further constrains the learned hashing functions so that they can locally adapt to the data structure of each modality, and Sparse Multi Modal Hashing($\mathrm{SM}^2\mathrm{H}$) [Wu *et al.*, 2014] preserves both data similarity based on joint multi-modal dictionary learning constrained by Hypergraph Laplacian sparse coding. Multimodal Latent Binary Embedding(MLBE) [Zhen and Yeung, 2012b] derives a probabilistic model to learn cross-modal binary hash codes.

In this work, we propose a new cross-modal hashing method. We take both hashing function learning and quantization of hash codes into consideration. This is inspired by the significance of quantization in single-modal hashing algorithms (e.g. ITQ [Gong and Lazebnik, 2011]). However, the effect of hash code quantization is less studied in cross-modal hashing literature. In conventional cross-modal hashing methods, the quantization of hash codes is realized independent of learning the correlation across domains. The separation of the two processes usually results in a sub-optimal solution. In this work, we propose a joint learning schema that consolidates the minimization of quantization loss and maximization of domain correlation into a single objective function so that optimal cross-modal hash codes are derived. In this way, we can simultaneously optimize the quantizer along with binary code learning, so that the quantizer is more fit to the cross-modality data search. To the best of our knowledge, it is the first attempt to integrate hash function learning with quantization together for cross-modal hashing. We call the proposed cross-modal hashing algorithm as *Quantized Correlation Hashing* (QCH). Compared to existing works, the novelty of our work includes:

- A new cross-modal hashing model is optimized simultaneously on both hashing code learning and binary quan-

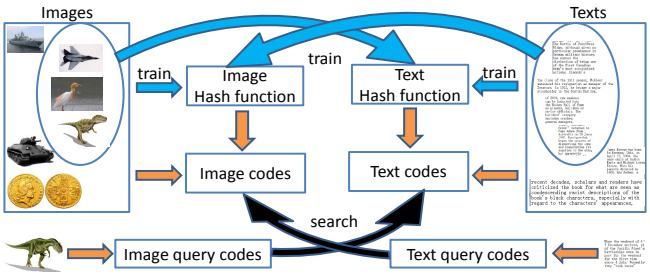


Figure 1: The framework of cross-modal hashing algorithm

tization;

- The proposed multi-modality objective function is transformed to a single-modality formalization, leading to an easier optimization procedure.

In the rest part of the paper, we present the details of the proposed method in Sec. 2 and extensive evaluation on three public datasets in Sec. 3. Finally, we conclude the paper in Sec. 4.

2 Quantized Correlation Hashing

In this section, we will present our QCH in details and use two-modalities for example introduction. Our method would be easily extended to the case for more than two modalities. In addition, our hashing model can be applied to deal with multi-view data as well, which will be tested in Section 3.

Assume that there are n objects represented by two modalities $(\mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{x}_i^T$ is the i^{th} row of data matrix $\mathbf{X} \in \mathbf{R}^{n\times d_x}$ of the first modality and $\mathbf{y}_i^T$ is the i^{th} row of data matrix $\mathbf{Y} \in \mathbf{R}^{n\times d_y}$ of the second modality. d_x and d_y are the dimensions of the two modalities. We assume all data are zero-centered which is usually adopted in existing hashing algorithms [Wang *et al.*, 2010; Zhang and Li, 2014], i.e., $\sum_{i=1}^{n}\mathbf{x}_i = \mathbf{0}$ and $\sum_{j=1}^{n}\mathbf{y}_j = \mathbf{0}$. In addition, the similarity information between data points across domains is given: $\mathbf{S}_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{y}_j$ are similar and $\mathbf{S}_{ij} = 0$ otherwise.We focus on a weakly supervised case which we are offered only positive pairs data.

The goal of cross-modal hashing is to learn two types of binary codes ($\mathbf{B_x} \in \{-1,1\}^{n\times c}$ and $\mathbf{B_y} \in \{-1,1\}^{n\times c}$) with the same code length c for each object by the hash functions $f(\mathbf{x}_i) = sign(\mathbf{W_x}^T\mathbf{x}_i)$ and $g(\mathbf{y}_j) = sign(\mathbf{W_y}^T\mathbf{y}_j)$, where $\mathbf{W_x} \in \mathbf{R}^{d_x\times c}$ and $\mathbf{W_y} \in \mathbf{R}^{d_y\times c}$ denote two projection matrices for the two modalities, respectively. The cross-modality fast searching task includes 1) using texts to search for the related images, and 2) using images to query the related texts. In cross-modal hashing framework, these two tasks are translated below: 1) using the hash codes of texts to find the related images in the hash code set of images according to the hamming distance, and 2) using the hash codes of images to query texts within the hash code set of texts. The framework of our approach is shown in Figure 1.

The principle of cross-modal hashing is that the codes $\mathbf{B_x}$ and $\mathbf{B_y}$ of data points from the same class but from different modalities should be as similar as possible, while they should be as distinct as possible if they are from different classes.

2.1 Formulation

We use the cosine similarity between hash codes of two across-modality data points,

$$\cos(f(\mathbf{x}_i), g(\mathbf{y}_j)) = \frac{f(\mathbf{x}_i)^T g(\mathbf{y}_j)}{f(\|\mathbf{x}_i)\|_2 \|g(\mathbf{y}_j)\|_2} \tag{1}$$

to measure the similarity in the hash space. The cosine similarity between two data points should be as small as possible if they are from the same class and as large as possible if not. We use the projection vector to replace the hash codes and obtain the approximate cosine similarity:

$$\cos(f(\mathbf{x}_i), g(\mathbf{y}_j)) \approx \frac{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_y}^T\mathbf{y}_j}{\sqrt{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i}\sqrt{\mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j}}. \tag{2}$$

We further borrow the maximum margin criterion idea from [Sharma *et al.*, 2012; Li *et al.*, 2003b] and replace the ratio operation with the subtraction operation:

$$\left(\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_y}^T\mathbf{y}_j - \sqrt{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i}\sqrt{\mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j}\right) \tag{3}$$

The quantization loss, inspired by ITQ [Gong and Lazebnik, 2011], is defined as $\|\mathbf{B} - \mathbf{XW}\|_F^2$. Combining the similarity constraint across domains and the quantization losses over each domain, we have the optimization problem:

$$\begin{aligned}
&\min O(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y}) \\
&= (\|\mathbf{B_x} - \mathbf{XW_x}\|_F^2 + \|\mathbf{B_y} - \mathbf{YW_y}\|_F^2) - \alpha' \sum_{(i,j)} \mathbf{S}_{ij} \\
&\left(\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_y}^T\mathbf{y}_j - \sqrt{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i}\sqrt{\mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j}\right) \\
&s.t. \mathbf{W_x}^T\mathbf{W_x} = \mathbf{I}_{c\times c} \\
&\mathbf{W_y}^T\mathbf{W_y} = \mathbf{I}_{c\times c}
\end{aligned} \tag{4}$$

Here α' is the control parameter to balance the quantization loss and the cosine similarity constraint. The constraints, $\mathbf{W_x}^T\mathbf{W_x} = \mathbf{I}_{c\times c}$ and $\mathbf{W_y}^T\mathbf{W_y} = \mathbf{I}_{c\times c}$, are used to make $\mathbf{W_x}$ and $\mathbf{W_y}$ be orthogonal projections.

2.2 Relaxation

We first derive an upper bound of the objective function $O(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y})$ in Eq.(4) use it as a substitute for the objective function, leading to a relatively easily-optimized problem. Then we transform the multi-modality formulation into a unimodal formulation.

Exploring the well-known inequality,

$$\begin{aligned}
&\frac{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i + \mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j}{2} \geq \\
&\sqrt{\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i}\sqrt{\mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j}
\end{aligned} \tag{5}$$

we simplify the third term on the right hand side of Eq.(4), resulting in

$$\sum_{(i,j)} \mathbf{S}_{ij}(\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_y}^T\mathbf{y}_j - \frac{1}{2}(\mathbf{x}_i^T\mathbf{W_x}\mathbf{W_x}^T\mathbf{x}_i + \mathbf{y}_j^T\mathbf{W_y}\mathbf{W_y}^T\mathbf{y}_j)) \tag{6}$$

write it in a matrix form,

$$tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{S}\mathbf{Y}\mathbf{W_y}) - \frac{1}{2}(tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{L_x}\mathbf{X}\mathbf{W_x}) + tr(\mathbf{W_y}^T\mathbf{Y}^T\mathbf{L_y}\mathbf{Y}\mathbf{W_y})) \tag{7}$$

where $\mathbf{L_x}$ and $\mathbf{L_y}$ are diagonal matrix and the diagonal elements are the row sum and column sum of $\mathbf{S}$.

Then, the objective function in Eq.(4) becomes

$$\begin{aligned} &O'(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y}) \\ =&(\|\mathbf{B_x} - \mathbf{X}\mathbf{W_x}\|_F^2 + \|\mathbf{B_y} - \mathbf{Y}\mathbf{W_y}\|_F^2) \\ &-2\alpha(tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{S}\mathbf{Y}\mathbf{W_y}) \\ &+tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{L_x}\mathbf{X}\mathbf{W_x}) + tr(\mathbf{W_y}^T\mathbf{Y}^T\mathbf{L_y}\mathbf{Y}\mathbf{W_y})) \end{aligned} \tag{8}$$

where $\alpha = \frac{1}{2}\alpha'$ is the control parameter to keep a balance between the hash function learning stage.

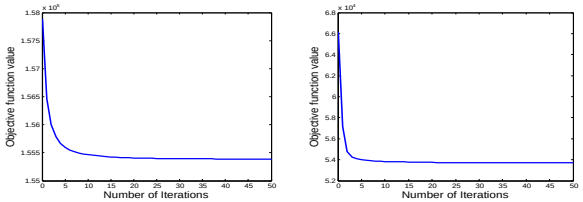
It can be distinctly verified that the relationship between the objective function values in Eq.(4) and Eq.(8) is: $O(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y}) \leq O'(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y})$. This indicates that $\min O'(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y}) \geq O(\mathbf{B_x}, \mathbf{B_y}, \mathbf{W_x}, \mathbf{W_y})$. In other words, our approach adopts a widely-used optimization principle: minimize an upper bound of the objective function.

To balance the cross-domain correlation $tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{S}\mathbf{Y}\mathbf{W_y})$ and its normalization term $tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{L_x}\mathbf{X}\mathbf{W_x}) + tr(\mathbf{W_y}^T\mathbf{Y}^T\mathbf{L_y}\mathbf{Y}\mathbf{W_y})$, we introduce an extra parameter β' to weigh the normalization term: $tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{S}\mathbf{Y}\mathbf{W_y}) - \frac{1}{2}\beta'(tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{L_x}\mathbf{X}\mathbf{W_x}) + tr(\mathbf{W_y}^T\mathbf{Y}^T\mathbf{L_y}\mathbf{Y}\mathbf{W_y})$. For simplicity, the third term and the fourth term in Eq.(8) are written as $2\alpha tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{S}\mathbf{Y}\mathbf{W_y}) - \beta(tr(\mathbf{W_x}^T\mathbf{X}^T\mathbf{L_x}\mathbf{X}\mathbf{W_x}) + tr(\mathbf{W_y}^T\mathbf{Y}^T\mathbf{L_y}\mathbf{Y}\mathbf{W_y})$, where $\beta = \alpha\beta'$.

Rather than optimizing the problem through an alternating procedure: optimizing the parameters for one domain and then optimizing the parameters for the other domain in one iteration, we transform the objective function by combining the parameters across domains together into a unimodal objective function. Define $\mathbf{W} = \begin{bmatrix} \mathbf{W_x} \\ \mathbf{W_y} \end{bmatrix}$, $\tilde{\mathbf{S}} = \begin{bmatrix} \beta\mathbf{L_x} & \alpha\mathbf{S} \\ \alpha\mathbf{S^T} & \beta\mathbf{L_y} \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{X} & \\ & \mathbf{Y} \end{bmatrix}$, and $\mathbf{B} = \begin{bmatrix} \mathbf{B_x} \\ \mathbf{B_y} \end{bmatrix}$. Then our objective function Eq.(8) can be simplified below:

$$O(\mathbf{B}, \mathbf{W}) = \|\mathbf{B} - \mathbf{Z}\mathbf{W}\|_F^2 - tr(\mathbf{W}^T\mathbf{Z}^T\tilde{\mathbf{S}}\mathbf{Z}\mathbf{W}). \tag{9}$$

To facilitate the optimization, we impose the orthogonality constraint over the whole weight matrix $\mathbf{W}$: $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{c\times c}$.



(a) NUS-WIDE (b) 58W-CIFAR

Figure 2: The objective value against the number of iterations on 2(a) NUS-WIDE and 2(b) 58W-CIFAR, respectively.

Algorithm 1 Quantized Correlation Hashing

Require:
Two-modality data matrices $\mathbf{X}$ and $\mathbf{Y}$, similarity matrix $\mathbf{S}$ between two modalities, code length c, random initialized matrix $\mathbf{W}^0$, α, β, iteration number N.
for $t = 1 : N$ **do**
  Optimize $\mathbf{B}$ when fixing $\mathbf{W}$ using Eq.(12);
  Iteratively optimize $\mathbf{W}$ when fixing $\mathbf{B}$ using Eq.(18);
end for
Ensure:
Projection matrix $\mathbf{W}$ and hash codes $\mathbf{B}$

2.3 Optimization

We adopt an alternating optimization procedure to iteratively optimize $\mathbf{W}$ and $\mathbf{B}$.

Fix W and optimize B

When $\mathbf{W}$ is fixed, the second term of Eq.(9) is a constant. Then, we have

$$\begin{aligned} O(\mathbf{B}) &= \|\mathbf{B}\|_F^2 + \|\mathbf{Z}\|_F^2 - 2tr(\mathbf{B}\mathbf{W}^T\mathbf{Z}^T) \\ &= nc + const - 2tr(\mathbf{B}\mathbf{W}^T\mathbf{Z}^T) \end{aligned} \tag{10}$$

Minimizing Eq.(10) is equal to maximizing the following:

$$tr(\mathbf{B}\mathbf{W}^T\mathbf{Z}^T) = tr(\mathbf{B}\mathbf{V}^T) = \sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{B}_{ij}\mathbf{V}_{ij} \tag{11}$$

where $\mathbf{V} = \mathbf{Z}\mathbf{W}$. Since $\mathbf{V}$ is fixed, it is clear that the hash codes $\mathbf{B}_{ij}$ should have the same sign as $\mathbf{V}_{ij}$. Consequently, we have

$$\mathbf{B} = sign(\mathbf{Z}\mathbf{W}) \tag{12}$$

Fix B and optimize W

When $\mathbf{B}$ is fixed, Eq.(9) becomes an optimization problem with orthogonal constraint. Through introducing Lagrangian multipliers, we can rewrite the objective function for optimizing $\mathbf{W}$ as follows:

$$L(\mathbf{W}, \mathbf{\Lambda}) = O(\mathbf{W}) - \frac{1}{2}tr(\mathbf{\Lambda}(\mathbf{W}^T\mathbf{W} - \mathbf{I})) \tag{13}$$

where $\mathbf{\Lambda}$ consists of Lagrangian multipliers, and $O(\mathbf{W})$ is

$$\begin{aligned} O(\mathbf{W}) &= \|\mathbf{B} - \mathbf{Z}\mathbf{W}\|_F^2 - tr(\mathbf{W}^T\mathbf{Z}^T\tilde{\mathbf{S}}\mathbf{Z}\mathbf{W}) \\ &= \|\mathbf{B}\|_F^2 + \|\mathbf{Z}\mathbf{W}\|_F^2 - 2tr(\mathbf{B}\mathbf{W}^T\mathbf{Z}^T) \\ &\quad - tr(\mathbf{W}^T\mathbf{Z}^T\tilde{\mathbf{S}}\mathbf{Z}\mathbf{W}) \\ &= const + tr(\mathbf{W}^T\mathbf{Z}^T\mathbf{Z}\mathbf{W}) - 2tr(\mathbf{B}\mathbf{W}^T\mathbf{Z}^T) \\ &\quad - tr(\mathbf{W}^T\mathbf{Z}^T\tilde{\mathbf{S}}\mathbf{Z}\mathbf{W}) \end{aligned} \tag{14}$$

Setting the gradient of Eq.(13) with respect to W to be zero, we can get

$$\frac{\partial L(\mathbf{W}, \mathbf{\Lambda})}{\partial \mathbf{W}} = \frac{\partial O(\mathbf{W})}{\partial \mathbf{W}} - \mathbf{W}\mathbf{\Lambda} = \mathbf{0} \tag{15}$$

For the convenience of description, let $\mathbf{G} = \frac{\partial O(\mathbf{W})}{\partial \mathbf{W}}$. Then we have

$$\mathbf{G} = \frac{\partial O(\mathbf{W})}{\partial \mathbf{W}} = 2(\mathbf{Z}^T\tilde{\mathbf{S}}\mathbf{Z}\mathbf{W} + \mathbf{Z}'^T\mathbf{Z}\mathbf{W} - \mathbf{Z}^T\mathbf{B}) \tag{16}$$

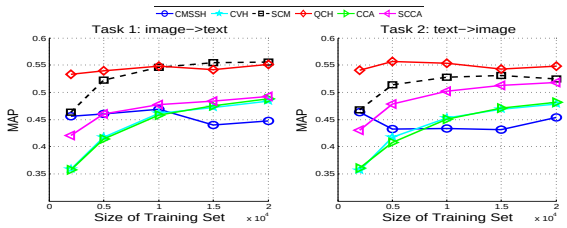


Figure 3: MAP results of different algorithms for two tasks on NUS-WIDE dataset with various numbers of training data.

From Eq.(15), it is clear that we can get $\mathbf{\Lambda} = \mathbf{W}^T\mathbf{G}$. Since $\mathbf{W}^T\mathbf{W}$ is symmetric, $\mathbf{\Lambda}$ is symmetric as well. So $\mathbf{\Lambda} = \mathbf{W}^T\mathbf{G} = \mathbf{G}^T\mathbf{W}$ and $\frac{\partial L(\mathbf{W})}{\partial \mathbf{W}} = \mathbf{G} - \mathbf{W}\mathbf{G}^T\mathbf{W}$. Based on the orthogonal constraint optimization procedure in [Wen and Yin, 2013], we can define a skew-symmetric matrix $\mathbf{A} = \mathbf{G}\mathbf{W}^T - \mathbf{G}^T\mathbf{W}$. Then, we will update $\mathbf{W}$ by Crank-Nicolson-like scheme [Smith, 1965]

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \frac{\tau}{2}\mathbf{A}(\mathbf{W}^{(t)} + \mathbf{W}^{(t+1)}) \tag{17}$$

where τ is the step size. By solving (17), we can obtain

$$\mathbf{W}^{(t+1)} = \mathbf{Q}\mathbf{W}^{(t)} \tag{18}$$

$$\mathbf{Q} = (\mathbf{I} + \frac{\tau}{2}\mathbf{A})^{-1}(\mathbf{I} - \frac{\tau}{2}\mathbf{A}) \tag{19}$$

Hereafter, we iteratively update $\mathbf{W}$ several times based on Eq.(18) with Barzilai-Borwein (BB) method [Wen and Yin, 2013]. In addition, please note that when iteratively optimizing $\mathbf{W}$, the initial $\mathbf{W}$ is set to be the one optimized in the last round between $\mathbf{B}$ and $\mathbf{W}$. For the first round, $\mathbf{W}$ is randomly initialized.

Convergence Analysis

$\mathbf{B}$ and $\mathbf{W}$ are alternately optimized for several iterations to seek an optimal solution. Since we minimize the objective function in each step, the convergence analysis of our optimization is

$$O(\mathbf{B}^{(t)}, \mathbf{W}^{(t)}) \geq O(\mathbf{B}^{(t+1)}, \mathbf{W}^{(t)}) \geq O(\mathbf{B}^{(t+1)}, \mathbf{W}^{(t+1)}) \tag{20}$$

where $\mathbf{B}^{(t)}$ and $\mathbf{W}^{(t)}$ are the optimal hash codes and projection matrix in the t^{th} iteration, respectively. In the experiments, the proposed hashing model almost converges at 50 iterations were conducted on optimizing $\mathbf{W}$, as shown in Figure 2(a) and 2(b). In summary, the whole procedure of our method is illustrated in Algorithm 1.

3 Experiments

3.1 Datasets

To verify the efficiency and effectiveness of QCH, a series of experiments are carried out on two benchmark multi-modal datasets, Wiki[Rasiwasia *et al.*, 2010] and NUS-WIDE [Chua *et al.*, 2009], and a large-scale dataset 58W-CIFAR [Krizhevsky and Hinton, 2009] for which we extracted two types of features to build multi-view data, so that cross-view retrieval can be performed.

The Wiki dataset [Rasiwasia *et al.*, 2010] consists of 2,866 documents containing image-text pairs annotated with 10 semantic labels and each image was represented by 128-dimensional SIFT feature and each text was denoted with a 10-dimensional feature vector generated by Latent Dirichlet Allocation (LDA) model.

The NUS-WIDE dataset [Chua *et al.*, 2009] consists of 269,648 images from 81 ground-truth concepts with a total number of 5,018 unique tags. In our experiments, 186643 samples from 10 classes that involve largest amount of data were selected. Besides, each image was presented by a 500-dimensional bag-of-words (BOW) feature vector and each text was denoted by a 1000-dimension tag vector.

To evaluate the performance in cross-view retrieval and the scalability of QCH, the 58W-CIFAR dataset consisting of 580,409 images from 10 categories selected from the 80 million tiny image dataset [Krizhevsky and Hinton, 2009] was used. To establish the multi-view data, a 384-dimensional GIST descriptor and a 496-dimensional HOG descriptor were extracted from each image. In the rest of paper, for the convenience of description, we consider GIST as "image" and HOG as "text" so that identical denotations can be acquired for all datasets.

On training and testing protocol, for Wiki dataset, 80% of the data were randomly selected as the training set and the remaining formed the testing set. For the other two datasets, NUS-WIDE and 58W-CIFAR, we randomly selected 1% of the data as the testing samples; while for training, we selected different numbers of instances to evaluate the influence of training size on our proposed model. In Experiment, we set the size of training set to 5000.

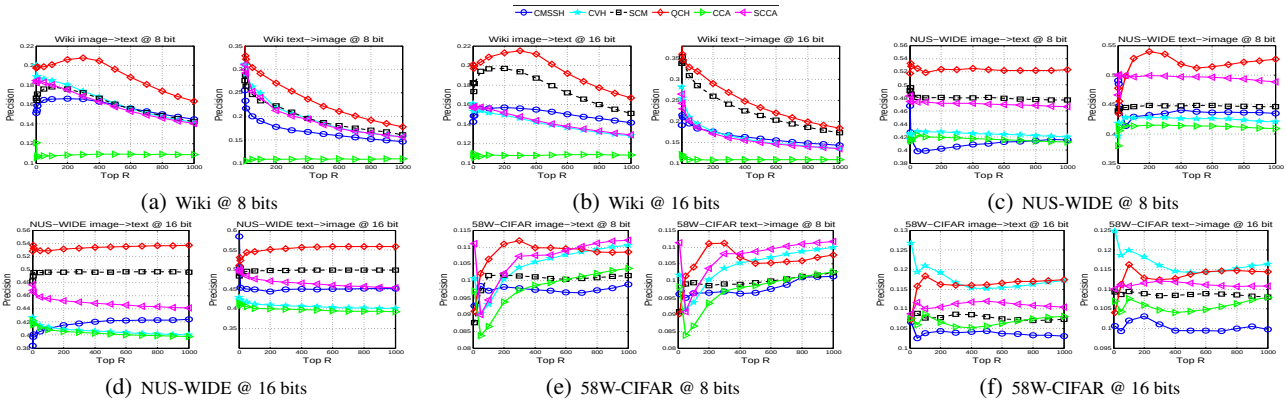
3.2 Compared Methods and Evaluation

In this experiment, we concentrate on two-modality data, and the following two tasks for fast search are conducted. 1) Task 1: using images to query texts; and 2) Task 2: using texts to query images. To make comparisons with QCH, three state-of-the-art cross-modal hashing algorithms were selected in this paper: (1) CMSSH [Bronstein *et al.*, 2010], (2) CVH[Kumar and Udupa, 2011] and (3) SCM [Zhang and Li, 2014]. CMSSH is a method to learn a group of hash functions for two modalities through eigen-decomposition and boosting. CVH extends spectral hashing [Weiss *et al.*, 2009] to multimodal data, and SCM is derived from Kernel-based supervised hashing [Liu *et al.*, 2012].

In addition, two baseline methods are selected, including (1) CCA [Hardoon *et al.*, 2004] which is an unsupervised and also based on the cosine similarity function that can be applied to match data points of two domains, and (2) supervised CCA(SCCA) which is our QCH method without incorporating the quantization loss for optimization.

To assess the performance of different algorithms, the widely used criterion Mean Average Precision(MAP)[Kumar and Udupa, 2011; Zhang and Li, 2014; Zhai *et al.*, 2013; Zhen and Yeung, 2012a] is selected, defined as:

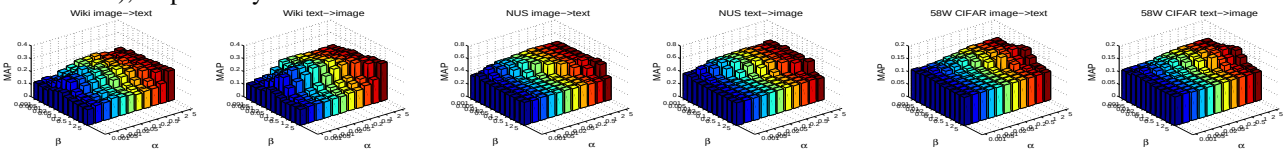
$$mAP = \frac{1}{Q}\sum_{i=1}^{Q} AP(\mathbf{q_i}) \tag{21}$$



(a) Wiki @ 8 bits (b) Wiki @ 16 bits (c) NUS-WIDE @ 8 bits

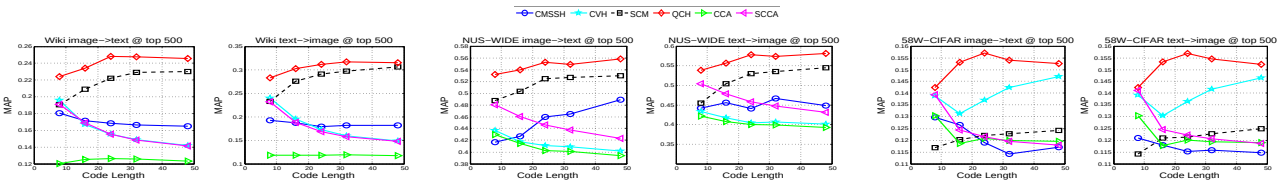
(d) NUS-WIDE @ 16 bits (e) 58W-CIFAR @ 8 bits (f) 58W-CIFAR @ 16 bits

Figure 4: Precision results of different algorithms for two taks with R varying from 50 to 1000 on the condition of two kinds of bit lengths: 8 bits (4(a) Wiki, 4(c) NUS-WIDE, and 4(e) 58W-CIFAR) and 16 bits (4(b) Wiki, 4(d) NUS-WIDE, and 4(f) 58W-CIFAR), respectively.



(a) Wiki (b) NUS-WIDE (c) 58W-CIFAR

Figure 5: Effect of parameters for QCH on different datasets (5(a) Wiki, 5(b) NUS-WIDE, and 5(c) 58W-CIFAR)



(a) Wiki (b) NUS-WIDE (c) 58W-CIFAR

Figure 6: MAP results of different algorithms with various code lengths for two tasks on the three datasets (6(a) Wiki, 6(b) NUS-WIDE, and 6(c) 58W-CIFAR).

where Q is the number of queries, and AP is computed as

$$AP(\mathbf{q}) = \frac{1}{L}\sum_{r=1}^{R} P_{\mathbf{q}}(r)\delta(r), \tag{22}$$

where L is the number of true neighbors for the query $\mathbf{q}$ among the retrieval items, $P_{\mathbf{q}}(r)$ denotes the precision for query point $\mathbf{q}$ when top r data points are returned, and $\delta(i)$ is an indicator function which is 1 when the ith result is a true neighbor of the query and otherwise 0. Ground-truth neighbors are defined as those pairs which share at least one label. In our experiments, we set $R = 500$. All results were averaged over 10 independent runs with random selection of training and testing data in each run. All the experiments were conducted on a workstation with 24 Intel(R) Xeon(R) E5-2620@2.0GHz CPUs, 96 GB RAM and 64-bit Ubuntu system.

3.3 Evaluation of QCH

Firstly, we investigate the influence of two parameters introduced in QCH: α and β. α controls the tradeoff between hash function learning stage and quantization stage and β is a regularizer coefficient. During this experiment, $c = 16$ is used. Fig.3 displays the MAP results of QCH on three datasets with α varying from 0.001 to 5 and β ranging from 0.001 to 5.

We can find that QCH is a little more sensitive to α and β on Wiki and NUS-WIDE as compared to the case on dataset 58W-CIFAR. By further investigating the results in this figure closely, an interesting and promising phenomenon can be found that when $\alpha > 0.02$ and $0.005 < \beta < 0.05$ on Wiki and NUS-WIDE datasets, QCH seems not so sensitive to α and β, which provides us the evidence that setting $\alpha = 0.05$ and $\beta = 0.02$ is a reasonable to QCH for Wiki and NUS-WIDE datasets. QCH performs better on 58W-CIFAR dataset when α and β are larger. Since multi-view data have stronger correlation than cross-modal data, so we set larger values, i.e. $\alpha = 1$ and $\beta = 0.1$ on multi-view datasets for example 58W-CIFAR.

3.4 Comparison with State-of-the-art Methods and Baselines

The comparison experiments were carried out between the compared methods and QCH on the three datasets. For fair comparison, the parameter settings of three state-of-the-art algorithms are adopted as recommended in their corresponding papers. Table 1 shows the experimental results of differen-

Table 1: MAP results of different algorithms with different code lengths on three datasets for different tasks. The best value along with each code length is highlighted.

Task	Method	Wiki					NUS-WIDE					58W-CIFAR				
		c = 8	c = 16	c = 24	c = 32	c = 48	c = 8	c = 16	c = 24	c = 32	c = 48	c = 8	c = 16	c = 24	c = 32	c = 48
Image Query v.s. Text database	CMSSH	0.1805	0.1716	0.1684	0.1663	0.1651	0.4171	0.4268	0.4579	0.4646	0.4887	0.1297	0.1265	0.1190	0.1142	0.1171
	CVH	0.1962	0.1671	0.1553	0.1487	0.1417	0.4366	0.4178	0.4112	0.4086	0.4015	0.1388	0.1313	0.1370	0.1424	0.1470
	SCM	0.1905	0.2089	0.2223	0.2289	0.2301	0.4876	0.5031	0.5246	0.5267	0.5300	0.1169	0.1204	0.1220	0.1236	0.1242
	CCA	0.1202	0.1253	0.1261	0.1257	0.1231	0.4293	0.4145	0.4023	0.4009	0.3939	0.1304	0.1187	0.1208	0.1198	0.1196
	SCCA	0.1910	0.1689	0.1552	0.1485	0.1412	0.4804	0.4601	0.4462	0.4372	0.4234	0.1394	0.1244	0.1215	0.1195	0.1180
	QCH	**0.2239**	**0.2343**	**0.2482**	**0.2477**	**0.2455**	**0.5319**	**0.5395**	**0.5528**	**0.5489**	**0.5584**	**0.1424**	**0.1530**	**0.1571**	**0.1548**	**0.1510**
Text Query v.s. Image database	CMSSH	0.1924	0.1874	0.1791	0.1819	0.1822	0.4416	0.4563	0.4411	0.4669	0.4482	0.1210	0.1180	0.1152	0.1158	0.1148
	CVH	0.2399	0.1955	0.1727	0.1592	0.1484	0.4366	0.4178	0.4112	0.4070	0.4008	0.1392	0.1304	0.1366	0.1417	0.1464
	SCM	0.2330	0.2762	0.2913	0.2979	0.3062	0.4547	0.5038	0.5298	0.5348	0.5443	0.1141	0.1209	0.1214	0.1227	0.1250
	CCA	0.1181	0.1180	0.1184	0.1193	0.1175	0.4215	0.4076	0.3998	0.3992	0.3930	0.1304	0.1178	0.1201	0.1193	0.1190
	SCCA	0.2326	0.1880	0.1685	0.1574	0.1480	0.5045	0.4780	0.4577	0.4470	0.4316	0.1411	0.1245	0.1222	0.1207	0.1188
	QCH	**0.2835**	**0.3034**	**0.3120**	**0.3170**	**0.3156**	**0.5386**	**0.5568**	**0.5776**	**0.5741**	**0.5814**	**0.1424**	**0.1533**	**0.1568**	**0.1556**	**0.1511**

t algorithms on two tasks with different code lengths on the three datasets. In addition, for different tasks, Fig. 4 further presents the precision changes of different algorithms along with the number of retrieval results R with $c = 8$ and $c = 16$, respectively. Based on the table and figure, we have the following analysis.

Task 1: using images to query texts

From Table 1, we can find that QCH outperforms all the compared algorithms on all datasets over all code lengths. From Fig.4, we can find that QCH is notably superior to the five compared methods on Wiki and NUS-WIDE datasets; On 58W-CIFAR, QCH is particularly better than SCM, CMSSH and CCA, and obtains a bit better performance than CVH and SCCA. The above performance of QCH demonstrates the efficiency and effectiveness of QCH on the first task. The superiority of QCH over SCCA on all datasets over all code lengths substantiates the importance of simultaneously optimizing quantization loss in our cross-modal hashing model.

Task 2: using texts to query images

From Table 1, the same conclusion as that of the first task can be drawn. Compared to the first task, we find that QCH obtains better performance on the second task, especially on Wiki and NUS-WIDE datasets. From Fig.4, we can observe that QCH shows its great advantage over all the compared algorithms on Wiki and NUS-WIDE. While on the 58W-CIFAR dataset, QCH achieves a notable improvement over SCM, CMSSH and CCA, and performs comparably to CVH and SCCA. Again, the comparison results on the second task consistently verify the superiority of the proposed cross-modal hashing model.

3.5 Effect of the Size of the Training Set and the Code Length

To further demonstrate the effectiveness of QCH, we additionally evaluate the effect of the code length and the effect of the numbers of training data against the compared methods in Figure 6 and Figure 3, respectively.

From Figure 6, we can conclude that QCH consistently performs significantly better than all compared algorithms on all datasets as the code length c increases no matter for Task1 or Task2. More specifically, QCH outperforms CMSSH, CVH, CCA and SCCA clearly; Compared to SCM, though QCH is a little better on Wiki and NUS-WIDE datasets, it particularly gains great advantage on 58W-CIFAR.

To see the effect of the training data size on QCH, we conducted experiments on the NUS-WIDE dataset by increasing the train size from 2,000 to 20,000 as shown in Figure 3. Evidently, QCH is not sensitive to the size, while the others are particularly sensitive. It indicates QCH can perform very well even only a small number of data points are used for training. This is beneficial from the incorporating of simultaneously optimizing quantization loss and optimizing similarity loss in the proposed model. Note that due to space limitation, we only report the results over the NUS-WIDE dataset, and the conclusions for other datasets hold.

4 Conclusions

In this paper, we introduce the Quantized Correlation Hashing (QCH) method for cross-modal similarity search that simultaneously optimizes the quantization loss and binary code learning. The problem is effectively optimized using the relaxation scheme and the alternative procedure. Extensive experiments on three datasets show that QCH outperforms the state-of-the-art cross-/multi-modal hashing methods, especially when the code length is small.

5 Acknowledgement

This research work was partially supported by Grants 973-2015CB351800, Natural Science Foundation Of China (No. 61472456,61421062), Guangzhou Pearl River Science and Technology Rising Star Project under Grant 2013J2200068, and the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014265. This work was also supported in part by Guangdong Provincial Government of China through the Computational Science Innovative Research Team Program.

References

[Bronstein *et al.*, 2010] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Da-

ta fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, 2010.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*, 2009.

[Datar *et al.*, 2004] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SOCG*, 2004.

[Gong and Lazebnik, 2011] Yunchao Gong and Svetlana Lazebnik. Iterative quantization a procrustean approach to learning binary codes. In *CVPR*, 2011.

[Gong *et al.*, 2014] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

[Hardoon *et al.*, 2004] David Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[Hwang and Grauman, 2012] Sung Ju Hwang and Kristen Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153, 2012.

[Kim and Choi, 2013] Saehoon Kim and Seungjin Choi. Multi-view anchor graph hashing. In *ICASSP*, 2013.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *CSD*, 2009.

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

[Li *et al.*, 2003a] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. Multimedia content processing through cross-modal association. In *ACM MM*, 2003.

[Li *et al.*, 2003b] Haifeng Li, Tao Jiang, and Keshu Zhang. Efficient and robust feature extraction by maximum margin criterion. In *NIPS*, pages 97–104, 2003.

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

[Masci *et al.*, 2013] Jonathan Masci, M Bronstein, A Bronstein, and Jürgen Schmidhuber. Multimodal similarity-preserving hashing. *TPAMI*, 2013.

[Norouzi and Blei, 2011] Mohammad Norouzi and David M Blei. Minimal loss hashing for compact binary codes. In *ICML*, 2011.

[Quadrianto and Lampert, 2011] Novi Quadrianto and Christoph H Lampert. Learning multi-view neighborhood preserving projections. In *ICML*, 2011.

[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 *ACM MM*, 2010.

[Sharma *et al.*, 2012] Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*. IEEE, 2012.

[Smith, 1965] Gordon D Smith. Numerical solution of partial differential equations. 1965.

[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 *SIGMOD*, 2013.

[Wang *et al.*, 2010] Jun Wang, Sanjiv Kumar, and Shih-Fu Chang. Semi-supervised hashing for scalable image retrieval. In *CVPR*, 2010.

[Weiss *et al.*, 2009] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In *NIPS*, 2009.

[Wen and Yin, 2013] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[Wu *et al.*, 2014] Fei Wu, Zhou Yu, Yi Yang, Siliang Tang, Yin Zhang, and Yueting Zhuang. Sparse multi modal hashing. *T MultiMedia*, 2014.

[Zhai *et al.*, 2013] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, and Wen Gao. Parametric local multimodal hashing for cross-view similarity search. In *IJCAI*, 2013.

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

[Zhang *et al.*, 2011] Dan Zhang, Fei Wang, and Luo Si. Composite hashing with multiple information sources. In *SIGIR*, 2011.

[Zhen and Yeung, 2012a] Yi Zhen and Dit-Yan Yeung. Co-regularized hashing for multimodal data. In *NIPS*, 2012.

[Zhen and Yeung, 2012b] Yi Zhen and Dit-Yan Yeung. A probabilistic model for multimodal hash function learning. In *SIGKDD*, 2012.

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