Discriminative Coupled Dictionary Hashing for Fast Cross-Media Retrieval

Zhou Yu
College of Computer Science
Zhejiang University, China
yuz@zju.edu.cn

Fei Wu
College of Computer Science
Zhejiang University, China
wufei@zju.edu.cn

Yi Yang
School of ITEE
The University of Queensland, Australia
yi.yang@uq.edu.au

Qi Tian
Dept. of Computer Science
University of Texas, USA
qitian@cs.utsa.edu

Jiebo Luo
Dept. of Computer Science
University of Rochester, USA
jluo@cs.rochester.edu

Yueting Zhuang
College of Computer Science
Zhejiang University, China
yzhuang@zju.edu.cn

ABSTRACT

Cross-media hashing, which conducts cross-media retrieval by embedding data from different modalities into a common low-dimensional Hamming space, has attracted intensive attention in recent years. The existing cross-media hashing approaches only aim at learning hash functions to preserve the intra-modality and inter-modality correlations, but do not directly capture the underlying semantic information of the multi-modal data. We propose a discriminative coupled dictionary hashing (DCDH) method in this paper. In DCDH, the coupled dictionary for each modality is learned with side information (e.g., categories). As a result, the coupled dictionaries not only preserve the intra-similarity and inter-correlation among multi-modal data, but also contain dictionary atoms that are semantically discriminative (i.e., the data from the same category is reconstructed by the similar dictionary atoms). To perform fast cross-media retrieval, we learn hash functions which map data from the dictionary space to a low-dimensional Hamming space. Besides, we conjecture that a balanced representation is crucial in cross-media retrieval. We introduce multi-view features on the relatively "weak" modalities into DCDH and extend it to multi-view DCDH (MV-DCDH) in order to enhance their representation capability. The experiments on two real-world data sets show that our DCDH and MV-DCDH outperform the state-of-the-art methods significantly on cross-media retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609563.

Keywords

Coupled dictionary learning; Cross-media retrieval; Hashing

1. INTRODUCTION

With the rapid development of Internet and social network, it has attracted increasing attention to study the correlations among multi-modal data. For example, an uploaded image on the Flickr web site is always tagged with some related descriptions or labels; a microblog may consist of a short text and correlative images. The relevant data from different modalities may have semantic correlations. Therefore, it is desirable to support cross-media retrieval across the data of different modalities, e.g., the retrieval of semantically-related textual documents in response to a query image and vice versa. Due to the large-scale nature of the existing multimedia data over the Internet, efficient retrieval of cross-media is particularly important.

An effective way to speed up the similarity search is the hashing-based method, which makes a tradeoff between accuracy and efficiency by *approximate* nearest neighbor search. The principle of hashing method is to map the high dimensional data into compact hash codes and generate the same or similar hash codes for similar data.

The motivation of hashing is to solve the approximate nearest neighbor (ANN) search problem. However, in the cross-media retrieval, the NN cannot be directly obtained as the data may come from different modalities. Therefore, most of the existing hashing approaches are not applicable to cross-media retrieval and cross-media hashing method should be specifically studied.

Generally speaking, the existing hashing approaches can be classified into three categories:

- **uni-modal hashing:** uni-modal hashing utilizes only a single type of feature (*homogeneous feature*) from uni-modal data as input, aiming at learning hash functions to project the homogeneous feature to compact hash codes.
- **multi-view hashing:** multi-view hashing utilizes multiple types of features (*heterogeneous features*) from uni-modal data as input, and learns hash functions to project the heterogenous features to hash codes.

- **cross-media hashing:** cross-media hashing utilizes data from multi-modalities (e.g, images and texts) as input, and preserves the intra-modality similarity and inter-modality correlation to learn hash functions. Thus, the correlation of the data from different modalities is measurable.

Most of the existing hashing approaches are uni-modal hashing. One of the well-known uni-modal hashing method is Locality Sensitive Hashing (LSH) [2], which uses random projections to obtain the hash functions. However, due to the limitation of random projection, LSH usually needs a quite long hash code and hundreds of hash tables to guarantee good retrieval performance. To make the hash codes compact, several learning based approaches are proposed. Weiss *et al.* proposed Spectral Hashing (SH) [23] which utilizes the distribution of training data and uses eigenfunction to obtain the hash functions. Compared with LSH, SH achieves better performance since the hash functions capture the manifold structure of the data. Since then, many extensions of SH have been proposed [27, 11, 20, 21, 5, 12].

However, in the real world applications, we can extract heterogenous features from the data and some multi-view hashing approaches are therefore leveraged to boost the retrieval performance[26, 17]. The principal idea of them is to learn the hash functions while preserving the local structures of each individual feature and globally considering the consistency of multi-view features.

Cross-media hashing is a new research area and there has been only limited research efforts focusing on it so far [3, 9, 28, 29, 15, 30, 25]. Most of the existing cross-media hashing approaches share the common idea of learning different hash functions individually for each modality and map the data from different modalities to a shared low-dimensional Hamming space. However, such a binary embedding strategy often results in poor indexing performance for the shared embedding space is not semantically discriminative, which is significantly important for cross-media retrieval.

In this paper, we propose a cross-media hashing framework titled Discriminative Coupled Dictionary Hashing (DCDH). Firstly, data from different modalities along with their *classes* or *categories* are jointly utilized to learn the both *discriminative* and *coupled* dictionaries. The discriminative capability indicates that data from same category will have similar sparse representation (i.e., sparse codes), and the coupling means not only intra-modality similarity but also inter-modality correlation will be preserved. As a result, DCDH assigns an explicit semantic meaning (i.e., topic) to each dictionary atom in multi-modal dictionaries and thus makes the sparse representation for the multi-modal data *interpretable*. Secondly, the obtained sparse codes for the data over their corresponding dictionary are exploited to learn the hash functions and further transform the sparse codes to compact binary hash codes.

Furthermore, we find that the representation capability of the dictionaries from different modalities varies and an "unbalanced" representation may adversely influence the performance of cross-media hashing. To address this problem, we additionally incorporate multi-view features into DCDH to enhance the representation capability of the dictionaries from the relatively "weak" modalities. This extended version of DCDH is named Multi-View DCDH (MV-DCDH).

The main contributions of this paper are three-fold:

- We propose a two-stage cross-media hashing framework consisting of the learning of discriminative coupled dictionaries and hash functions, respectively. The learned discriminative coupled dictionaries in the first stage have both discriminative and similarity-preserving capability.
- The discriminative coupled dictionary learning is formulated as an optimization problem of submodular function and an approximation solution can be efficiently obtained using a greedy algorithm.
- Multi-view and multi-modal data are jointly considered in the MV-DCDH framework. The multi-view features is incorporated to strengthen the representation capability for the dictionary from a relatively "weak" modality and lead to a balanced cross-media representation. This enhancement improves the cross-media retrieval performance of DCDH significantly.

The rest of the paper is organized as follows: In Section 2, we review the related work of dictionary learning and cross-media hashing approaches. In Section 3, we give out the detailed explaination of our DCDH and its multi-view extension MV-DCDH. The complexity of DCDH is analyzed in Section 4. Experimental results and comparisons on two real-world data sets are demonstrated in Section 5. Finally, the conclusions are given.

2. RELATED WORK

2.1 Dictionary Learning

Beyond the traditional dictionary learning approaches [24, 1], coupled or semi-coupled dictionary learning approaches [6, 22] attempt to learn dictionaries for multi-modal data by minimizing the reconstruction error of each dictionary and preserving the *pairwise* correspondence across different modalities. However, these approaches are unsupervised so that the *class* or *category* information is not exploited and can not significantly boost the performance of learned coupled dictionaries. Zhuang *et al.* proposed a supervised semi-coupled dictionary learning approach which introduces the category side information into multi-modal dictionary learning via a $\ell_{2,1}$-norm regularization term.

Our proposed DCDH bears some resemblance to submodular dictionary learning (SDL) [7] that takes advantage of the submodularity to learn dictionary efficiently. We extend the idea from *uni-modal* data into *multi-modal* data in order to learn discriminative coupled dictionaries .

2.2 Cross-media Hashing

Cross-media retrieval is a hot research focus in recent years [16, 32, 31]. With the rapid advance of hashing, some cross-media hashing approaches have been proposed [3, 9, 28, 29, 15, 30, 25].

The problem of cross-media hashing was first proposed by Bronstein *et al.* in CMSSH [3]. Specifically, given two modalities of data sets, CMSSH learns two groups of hash functions to ensure that if two data points (with different modalities) are relevant, their corresponding hash codes are similar and otherwise dissimilar. However, CMSSH only preserves the inter-modality correlation but ignores the intra-modality similarity. Kumar *et al.* extended Spectral Hashing [23] from the traditional uni-modal setting to the

multi-modal scenario and proposed CVH [9]. CVH attempts to generate the hash codes by minimizing the distance of hash codes for the similar data and maximizing the distance for the dissimilar data. The inter-view and intra-view similarities are both preserved in CVH. LCMH [30] adopts a "two-stage" strategy to learn the cross-media hash functions: First, the data within each modality are low-rank represented using the anchor graph[11]. Then, hash functions for each modality are learned to project the data from each anchor graph space into a shared Hamming space. MLBE employs a probabilistic generative model to encode the intra-similarity and inter-similarity of data across multiple modalities. According to the estimation of maximum a posteriori, the binary latent factors can be obtained and then be taken as the hash codes in MLBE. However, the hash codes generated by MLBE do not require the independency between different hash bits, and may obtain highly redundant hash bits.

Wu *et al.* introduced dictionary learning into cross-media hashing [25]. By the joint modeling of the intra-modality similarity and inter-modality correlation among multi-modal data with a hypergraph, the coupled dictionaries with a hypergraph laplacian regularizer are learned in an iterative manner. The learned dictionary for each modality is then adopted as the hash functions, and the sparse code of each data point over its corresponding dictionary is regarded as the hash code to perform cross-media retrieval.

Inspired by the effectiveness of using the coupled dictionary space to represent the data points from different modalities, we additionally emphasize discrimination when learning coupled dictionaries in order to make the shared dictionary space interpretable. Furthermore, unlike [25] that directly exploits the sparse codes as the hash codes, we further learn hash functions to map the sparse codes to binary hash codes.

3. THE OVERVIEW OF DCDH

In this section, we introduce the detail of DCDH. Figure 1 illustrates the algorithmic flowchart of our DCDH. For the sake of illustrative simplicity, we assume that only two kinds of data (e.g., images and texts) are available in Figure 1. The proposed DCDH mainly consists of the following two stages:

1. *Discriminative coupled dictionary learning*: In Figure 1, the clusters of multi-modal data can be learned by a submodular function with the help of inter-modality correlation, intra-modality similarity as well as the supervised side information (e.g.,category labels). Given a cluster, its centroid is taken as a dictionary atom of their corresponding dictionary. That is to say, the dictionary atom from one modality is *coupled* with the corresponding dictionary atoms from the other modalities in the same cluster.

2. *Unified hash functions learning*: Based on the learned coupled dictionaries, the data points from different modalities can be represented as sparse codes in a unified dictionary space. Afterwards, utilizing the sparse property, hash functions which project the sparse codes into compact binary hash codes can be learned efficiently.

The notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbols	Explanation
M	the number of modalities
m	$m \in \{1, 2, ...M\}$ is one of the M modalities
K	the common size of the coupled dictionaries
L	the length of the hash codes
N	the common size of each data set
p^1,...,p^M	the dimensionality of each modality
X^1,...,X^M	data set $X^m = [x_1^m, x_2^m, ..., x_{N_m}^m] \in \mathbb{R}^{p^m \times N}$
D^1,...,D^M	dictionary $D^m = [d_1^m, d_2^m, ..., d_K^m] \in \mathbb{R}^{p^m \times K}$
Z^1,...,Z^M	sparse codes $Z^m \in \mathbb{R}^{K \times N}$ of X^m w.r.t. D^m

3.1 Unified Graph Representation of Labeled Multi-modal Data

To well model the intra-modality similarity and the inter-modality correlation of M data sets, we resort to the unified graph $G(V, E, w)$ similar to [19]. The vertex set V denotes the data from all the data sets, and the edge set E models the pairwise intra-modality similarity or the inter-modality correlation between data points. The weight of an edge is measured by some similarity functions which we will discuss in the following.

To model the intra-modality similarity within the same modality, we adopt the *local* similarity metric with a Gaussian kernel. The intra-modality similarity $w_{i,j}^m$ of two data points x_i^m and x_j^m from modality m is defined as:

$$w_{i,j}^m = \begin{cases} e^{-\frac{|x_i^m - x_j^m|^2}{2\sigma^2}}, & \text{if } x_i^m \in \mathcal{N}_k(x_j^m) \text{ or } x_i^m \in \mathcal{N}_k(x_j^m) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{N}_K(x)$ represents the set of k-nearest neighbours of x and $\sigma = \frac{1}{N}\sum_{i,j} |x_i^m - x_j^m|^2$ is the expectation over all the pairwise distance in X^m.

It takes $O(N^2 p^m)$ time to compute an intra-modality similarity matrix. When N is large, we can use some approximated methods such as the anchor graph structure to construct this similarity matrix efficiently [11]. In this paper, we simply use the exact k-NN graph in Eq.(1).

To model the inter-modality correlation of the data from two modalities (we name them as modality a and b, $a \neq b$ and $a, b \in \{1, 2, ..., M\}$), the similarity function $w_{i,j}^{a,b}$ for two data x_i^a and x_j^b is defined as:

$$w_{i,j}^{a,b} = \begin{cases} 1, & \text{if } x_i^a \text{ has known correlation with } x_j^b \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Moreover, to better understand the semantics of data, we additionally exploit the category information. Let C be the category-labels set indicating the category label of each data point (i.e., each vertex in G), the final category-labeled unified graph is denoted as $G(V, E, w, C)$.

3.2 Discriminative Coupled Dictionary Learning

Given the category-labeled graph $G(V, E, w, C)$, we attempt to jointly learn the *discriminative* coupled dictionaries D^1,...,D^M for the data from each modality. The coupling of the dictionaries indicates that these dictionaries have the same number of atoms (i.e, K) and the dictionary atoms from M modalities have a one-to-one correspondence

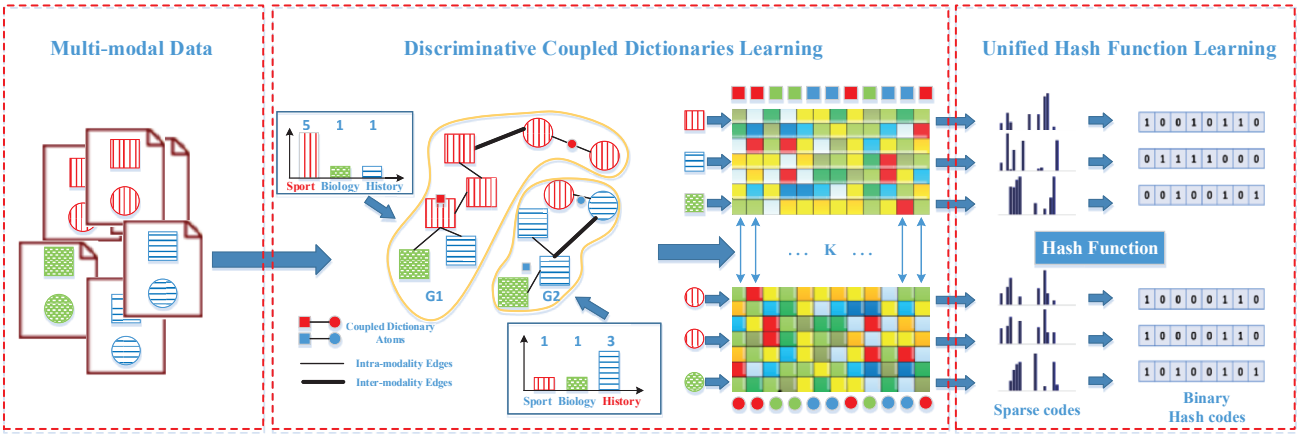


Figure 1: The algorithmic flowchart of DCDH. Without loss of generality, we assume that there are two modalities of data (represented as squares and circles). The data points with the same stripe have the same category label (e.g., 'sports', 'biology', 'history'). Given multi-modal data, the submodular dictionary learning is utilized to obtain discriminative coupled dictionaries. The "discriminative" capability is reflected by the fact that each dictionary atom is assigned a dominant category label to enhance its interpretability (i.e., category 'sport' for G1 and 'history' for G2). The "coupling" means that each dictionary atom in one modality has its counterpart dictionary atom in another modality. The coupled dictionary atoms are combined to characterize the multi-modal data. Each data point from a given modality can be sparsely represented as a sparse code using its corresponding dictionary. Finally, hash functions is learned to transform the sparse codes to binary hash codes.

(paired dictionary atoms), since the paired dictionary atoms have their different intrinsic power to characterize the multi-modal data. Moreover, the paired dictionary atoms are discriminative in terms of semantics (i.e., category) and is consistent with only one category label. That is to say, the data from different modalities are semantically aligned in a shared *coupled dictionary space*.

Inspired by the efficiency and effectiveness of submodular dictionary learning approach [7], we formulate our discriminative coupled dictionary learning as a graph partition problem on $G(V, E, w, C)$. Learning coupled dictionaries with size K is equal to partitioning the category-labeled graph G into K subgraphs which can be further regarded as a problem of selecting a subset A of the edge set E (i.e., $A \subseteq E$) [7, 10]. We can formulate an objective function with respect to A and maximize it to obtain the optimal partitions. Our objective function has the property of submodularity and thus can be approximately optimized with an efficient greedy algorithm.

Our objective function consists of three parts which corresponds to the following requirements: 1) each subgraph should be *compact* so that the obtained dictionaries have a good representative capability; 2) each subgraph is encouraged to be *discriminative* so that the sparse representation of the data over learned dictionary (i.e. using the centroid to represent subgraphs), from the same category to be similar; 3) to avoid the over-fitting on the subgraphs' size (some subgraphs may be extremely large while the others are so tiny), the size of each subgraph is in *balance* (nearly equal).

Compact Function: The entropy rate of the random walk over the graph G is exploited to obtain the compact subgraphs. The entropy rate measures the uncertainty of a stochastic process $S = \{S_t | t \in T\}$ where T is an index set. For a discrete random process, the entropy rate is defined as an asymptotic measure as: $\mathcal{H}(S) = \lim_{t\to\infty} H(S_t|S_{t-1}, ..., S_1)$, which is the conditional entropy of the last random variable given the past. In the case of a stationary 1st-order Markov chain, the entropy rate is: $\mathcal{H}(S) = \lim_{t\to\infty} H(S_t|S_{t-1}) = \lim_{t\to\infty} H(S_2|S_1) = H(S_2|S_1)$.

We define the random walk model on graph G as $S = \{S_t | t \in T\}$. The transition probability from the vertex v_i to the vertex v_j is defined as $p_{i,j} = Pr(S_{t+1} = v_j | S_t = v_i) = w_{i,j}/w_i$ where $w_i = \sum_{k:e_{i,k}\in E} w_{i,k}$ is the sum of incident weights of the vertex v_i, and the stationary distribution is defined as:

$$\mu = (\mu_1, \mu_2, ..., \mu_{|V|})^T = (\frac{w_1}{w_{all}}, \frac{w_2}{w_{all}}, ..., \frac{w_{|V|}}{w_{all}})^T \quad (3)$$

where $w_{all} = \sum_{i=1}^{|V|} w_i$ is the sum of incident weights of all vertices. The entropy rate of the random walk is defined as:

$$\begin{aligned} \mathcal{H}(\mathbf{S}) &= H(S_2|S_1) = \textstyle\sum_i \mu_i H(S_2|S_1 = v_i) \\ &= -\textstyle\sum_i \mu_i \sum_j P_{i,j} \log P_{i,j} \end{aligned} \quad (4)$$

Leaving μ in Eq.(3) intact, the set functions for the transition probability $P_{i,j} : 2^E \to R$ w.r.t. A are defined as:

$$P_{i,j}(A) = \begin{cases} \frac{w_{i,j}}{w_i}, & \text{if } i \neq j \text{ and } e_{i,j} \in A \\ 0, & \text{if } i \neq j \text{ and } e_{i,j} \notin A \\ 1 - \frac{\sum_{j:e_{i,j}\in A} w_{i,j}}{w_i}, & \text{if } i = j \end{cases} \quad (5)$$

Consequently, the compact function with respect to A can be defined as the entropy rate of the random walk on G:

$$\mathcal{H}(A) = -\sum_i \mu_i \sum_j P_{i,j}(A) \log P_{i,j}(A) \quad (6)$$

Given the entropies of the transition probabilities, maximizing the entropy rate in Eq.(6) encourages the edges with large weights (small distance) to be selected [7]. Hence the compact function $\mathcal{H}(A)$ can generate compact subgraphs.

Discriminative Function: To encourage the discrimination of subgraphs which further guarantees the sparse representation of the data from the same category to be similar, a discriminative function on G is proposed [7].

Let A be the selected edge set, N_A be the number of subgraphs with respect to A, the partition of graph G with selected edge set A is $\mathcal{G}_A = \{G_1, ..., G_{N_A}\}$ where each G_i is a subgraph. We construct a count matrix $\mathbf{N} = [\mathbf{N}^1, ..., \mathbf{N}^{N_A}] \in \mathbb{R}^{c \times N_A}$ for the count of each category label of the data assigned to each subgraph and c is the number of the categories of the multi-modal data set. Each $\mathbf{N}^i = [N_1^i, ..., N_c^i]^T \in \mathbb{R}^c$ where N_c^i is the number of data points from the c-th category assigned to i-th subgraph. It is worth noting that the size of the count matrix $\mathbf{N}$ is dynamic since N_A changes when new edges are added to the selected edge set A.

The purity for each subgraph G_i is defined as: $\mathcal{P}(G_i) = \frac{1}{C_i}\max_y N_y^i$ where $y \in \{1, 2, ..., c\}$ is the category label, $C_i = \sum_{y=1}^{c} N_y^i$ is the count for data points of all categories assigned to subgraph G_i. The overall purity of $\mathcal{G}_A$ is:

$$\mathcal{P}(\mathcal{G}_A) = \sum_{i=1}^{N_A} \frac{C_i}{C_{total}} \mathcal{P}(G_i) = \sum_{i=1}^{N_A} \frac{1}{C_{total}} \max_y N_y^i \quad (7)$$

where $C_{total} = \sum_i C_i = |V|$ is the sum of the count of all subgraphs. The discriminative function is defined as:

$$\mathcal{D}(A) = \mathcal{P}(\mathcal{G}_A) - N_A = \sum_{i=1}^{N_A} \frac{1}{C_{total}} \max_y N_y^i - N_A \quad (8)$$

$\mathcal{D}(A)$ measures the discriminative capability of the subgraphs. Maximizing $\mathcal{D}(A)$ encourages each subgraph to have a consistent category label, i.e., the data within each subgraph are expected to have the same category label.

Balancing Function: If we only use the compact and discriminative functions, there may exist some extreme cases where the majority of data belong to one subgraph and the other data are sporadically dispersed. This makes the learned dictionary suffer from over-fitting. Therefore, a balancing function is used to regularize the subgraphs of similar sizes.

Denote p_A as the distribution of the subgraph membership, p_A is formulated as:

$$p_A(i) = \frac{|G_i|}{\sum_i |G_i|}, \ i = \{1, 2, ..., N_A\} \quad (9)$$

The balancing function is defined using the entropy maximum theory:

$$\mathcal{B}(A) = -\sum_i p_A(i) \log(p_A(i)) - N_A \quad (10)$$

The aforementioned three functions are proved to be monotonically increasing and submodular with respect to A [10][7]. Furthermore, It has been proved that the linear combination of submodular function is still submodular [14]. Therefore, we define an overall function $\mathcal{F} = \mathcal{H}(A) + \lambda\mathcal{D}(A) + \gamma\mathcal{B}(A)$ which is also monotonically increasing and submodular. The optimal solution of $\mathcal{F}(A)$ is achieved by maximizing the objective function with best A as:

$$\begin{array}{ll} \max\limits_{A} & \mathcal{H}(A) + \lambda\mathcal{D}(A) + \gamma\mathcal{B}(A) \\ s.t. & A \subseteq E \text{ and } N_A \geq K \end{array} \quad (11)$$

where λ and γ control the contribution of the three terms. Follow the settings of [7], we set $\lambda = \frac{\max_{e_{i,j}} \mathcal{H}(e_{i,j}) - \mathcal{H}(\emptyset)}{\max_{e_{i,j}} \mathcal{D}(e_{i,j}) - \mathcal{D}(\emptyset)} \lambda'$ and $\gamma = \frac{\max_{e_{i,j}} \mathcal{H}(e_{i,j}) - \mathcal{H}(\emptyset)}{\max_{e_{i,j}} \mathcal{B}(e_{i,j}) - \mathcal{B}(\emptyset)} \gamma'$, where λ' and γ' are predefined parameters. $N_A \geq K$ is a constraint on the number of subgraphs which enforces exactly K subgraphs since the objective function is monotonically increasing.

Directly maximizing Eq.(11) is a NP-hard problem. However, since $\mathcal{F}(A)$ is a submodular function, we can obtain an approximate solution by a simple greedy algorithm, which gives a $\frac{1}{2}$-approximation lower bound on the optimality of the solution [14]. When the optimal K subgraphs are obtained, we simply use the *center* of the data within each subgraph as the corresponding dictionary atom. Since each subgraph consists of the data from M modalities respectively, M coupled dictionary atoms are obtained. The coupled dictionaries of the common size K are generated based on all the subgraphs. The overall algorithm of the discriminative coupled dictionary learning is summarized in Algorithm 1.

Note that the weights of the intra-modality edges $w_{i,j}^m \in (0, 1]$ are not larger than the inter-modality edges $w_{i,j}^{a,b} = 1$, and the two vertices connected by the inter-modality edges are within the same category. Therefore, adding an inter-modality edge satisfies the discriminative function and the compact function at the same time and each inter-modality edge has a high probability to be selected out at the early iterations in the step 4 of Algorithm 1. This observation is valuable which ensures that within each subgraph, there is at least one data point for every modality and the trivial zero-value dictionary atoms is avoided.

Algorithm 1 Discriminative Coupled Dictionary Learning

Input: data sets X^1,...,X^M, λ', γ', K
Output: The learned coupled dictionaries D^1,...,D^M

1: Construct unified graph with labeled multi-modal data $G(V, E, w, C)$ for the data sets X^1,...,X^M
2: Initialization: $A \leftarrow \emptyset$, $D^1, ..., D^M \leftarrow \emptyset$
3: **while** $N_A > K$ **do**
4: $\quad e^* = \operatorname{argmax}_{e \in E} \mathcal{F}(A \cup \{e\}) - \mathcal{F}(A)$
5: $\quad A \leftarrow A \cup \{e^*\}$
6: **end while**
generation of coupled dictionaries.
7: **for** each subgraph G_i in G **do**
8: $\quad$ **for** $m = 1$ to M **do**
9: $\quad\quad V_i^m = \{v_j | v_j \in G_i \text{ and } v_j \text{ from modality } m\}$
10: $\quad\quad D^m \leftarrow D^m \cup \{\frac{1}{|V_i^m|} \sum_{j: v_j \in V_i^m} v_j\}$
11: $\quad$ **end for**
12: **end for**

3.3 Unified Hash Function Learning

As the coupled dictionary for each modality is learned, the data from each modality can be encoded as sparse codes using its corresponding learned dictionary.

For the data points in data set X^m, their K-dimensional sparse codes Z^m can be efficiently computed using the dictionary D^m as follows:

$$\begin{array}{ll} \min\limits_{Z^m} & \|X^m - D^m Z^m\|_F^2 + \beta \|Z^m\|_1 \\ s.t. & Z^m \geq 0 \end{array} \quad (12)$$

The non-negative constraint on Z^m is needed for the following hash functions learning step. Eq.(12) is a simple non-negative LASSO problem [18], and we use the efficient LARS [4] solver to solve this problem. Moreover, despite of

different choices of β, the sparsity (maximum number of the zero-elements) of Z can be well controlled by LARS. This is helpful since we expect the sparsity of each sparse code to be equivalent. The sparsity is set to 0.9 (i.e., 90% elements of a sparse code are 0) throughout the paper.

By solving the Eq.(12) for the data set of each modality, the sparse codes $Z^1, ..., Z^M$ are correspondingly generated. Denote $Z = [Z^1, ..., Z^M] \in \mathbb{R}^{K \times MN}$ as the joint sparse codes for all M data sets , we intend to further learn hash functions which linearly projects each sparse code $z_i \in Z$ onto L-dimensional compact binary hash codes ($L < K$).

The commonly used hash function learning strategy is based on graph-laplacian [9, 27], etc. The hash functions are learned by solving an eigenvalue-decomposition problem of a laplacian matrix which takes $O(N^3)$ time. However, it is infeasible to learn hash functions when N is large. Therefore, we adopt the hash function learning strategy based on the *sparse* characteristic of sparse codes.

Since Z is non-negative, we can use Z (each column has been ℓ_1 normalized) to construct an approximate adjacency matrix $\hat{W} = Z^T \Lambda^{-1} Z \in \mathbb{R}^{N \times N}$ where $\Lambda = \text{diag}(Z\mathbf{1}) \in \mathbb{R}^{K \times K}$ [11]. The approximate adjacency matrix $\hat{W}$ is: 1) nonnegative and sparse; 2) low-rank (the rank is at most K); 3) double stochastic, i.e., has unit row and column sum. Afterwards, the laplacian matrix is formulated as $\hat{L} = I - \hat{W}$ where I is the identity matrix.

The optimal hash functions can be acquired as the L eigenvectors with smallest eigenvalues of the approximated laplacian matrix $\hat{L}$ (removing the trivial eigenvector corresponds to eigenvalue 0), which is equal to L eigenvectors with largest eigenvalues of $\hat{W}$. Due to the low-rank property of $\hat{W}$, a smaller matrix $Q = \Lambda^{-1/2} Z Z^T \Lambda^{-1/2} \in \mathbb{R}^{K \times K}$ is substituted for eigenvalue-decomposition problem on $\hat{L}$. By solving the eigen-system of Q, L largest eigenvector-eigenvalue pairs$\{(v_k, \sigma_k)\}_{k=1}^{L}$ where $1 > \sigma_1 \geq ...\sigma_L > 0$ are obtained. Denote $V = [v_1, v_2, ..., v_L] \in \mathbb{R}^{K \times L}$ and $\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_L) \in \mathbb{R}^{L \times L}$, the hash functions are defined as follows:

$$h(z) = \text{sign}(P^T z) \tag{13}$$

where $P = \sqrt{MN} \Lambda^{-1/2} V \Sigma^{-1/2} \in \mathbb{R}^{K \times L}$ is the normalized projection matrix, $z \in \mathbb{R}^K$ is a sparse code and $\text{sign}(\cdot)$ is the binary function.

When given a new data point, its hash code can be consequently generated using a *two-stage* mechanism: for example, given a data point x^m from modality m, it is first nonlinearly transformed to a sparse code z^m using its corresponding dictionary D^m similar with Eq.(12). After that, with the learned projection matrix P, z^m is linearly transformed to a L-dimensional compact binary hash code using the learned hash functions in Eq.(13).

3.4 Multi-View Enhancement

It is natural that the representation capability of the dictionaries for different modalities varies widely. For example, the representing capability of the dictionary for a text modality is much stronger than the one for an image modality. This "unbalanced" representation may lead to an unsatisfying cross-media retrieval performance. Therefore, we incorporate the multi-view representation into our coupled dictionary learning to enhance the representing capability of the relatively "weak" modalities.

Without loss of generality, assuming we have two modalities a and b, for modality a, we have a single-view feature X^a; for modality b, we extract multi-view (e.g., 2 views) features X^{b_1} and X^{b_2}. The construction of category-labeled unified graph $G(V, E, w, C)$ is similar with the aforementioned methods. The size of vertex set does not change since each vertex represents one data point. The edge set E is expanded as some relations between the vertices in V are added with the introduction of multi-view features.

The multi-view discriminative coupled dictionary learning method is similar to the DCDH in Algorithm (1) except for the generation of coupled dictionaries. For the single-view modality a, its corresponding dictionary D^a is learned; for the multi-view modality b, two dictionaries D^{b_1} and D^{b_2} are learned, respectively.

For modality a, we use the dictionary D^a to generate sparse codes Z^a for the data set X^a by Eq.(12). For modality b, we use the dictionaries D^{b_1} and D^{b_2} to jointly learn the sparse codes Z^b as follows:

$$\begin{aligned} \min_{Z^b} \quad & \sum_{i=1,2} \|X^{b_i} - D^{b_i} Z^b\|_F^2 + \beta \|Z^b\|_1 \\ & \Leftrightarrow \|X^b - D^b Z^b\|_F^2 + \beta \|Z^b\|_1 \\ s.t. \quad & Z^b \geq 0 \end{aligned} \tag{14}$$

where Z^b is the sparse codes over the multi-view dictionaries, $X^b = [X^{b_1}; X^{b_2}] \in \mathbb{R}^{(p^1+p^2) \times N}$, $D^b = [D^{b_1}; D^{b_2}] \in \mathbb{R}^{(p^1+p^2) \times K}$. Here, p^1, p^2 denote the dimensionality of the two views. The optimization of Eq.(14) is similar to Eq.(12).

4. COMPLEXITY ANALYSIS

Our DCDH approach consists of an off-line stage to learn the discriminative coupled dictionaries and unified hash functions; an on-line stage to encode an out-of-sample data point into a binary hash code. We detail the time complexity for each part respectively.

4.1 Off-line training

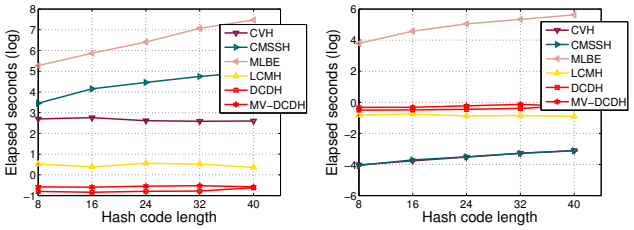
Discriminative coupled dictionary learning: Leaving out the time for constructing the graph $G(V, E, w, C)$ which takes $O(\sum_{m=1}^{M} N^2 p^m)$ time, the complexity of discriminative coupled dictionary learning can be implemented efficiently. Using a well designed heap structure, the ideal time complexity is $O(|V|\log|V|)$ (i.e., $O(MN\log MN)$) [7].

Unified hash functions learning: To learn the unified hash functions V, we first learn the sparse codes of $Z^1, ..., Z^M$ using Eq.(12). which can be solved in $O(\sum_{m=1}^{M} NKp^m)$ time using the LARS algorithm [4]. However, the generation of each sparse code is independent which can be solved in $O(p^m K)$ time, some parallel implementation can be adopted to solve the problem efficiently [1]. After the sparse codes for all training data are obtained, an eigen-system of a small matrix $Q \in \mathbb{R}^{K \times K}$ is solved in $O(K^3)$ time to obtain the projection matrix W and corresponding hash functions. Therefore, the overall unified hash functions learning step can be very efficient.

4.2 On-line hash encoding

The on-line hash encoding step should be fast enough to support the cross-media retrieval over the large scale data set. The time for encoding a new data point is two-fold:

[1] `http://spams-devel.gforge.inria.fr/`



(a) Off-line training time (b) On-line testing time

Figure 2: The time cost of the training and testing stages for DCDH, MV-DCDH and other cross-media hashing approaches. The experiments are conducted on Wiki-Potd data set with dictionary size K = 100 and MV-DCDH uses two views of features.

Sparse Coding: Given a new data point x^m from modality m, its sparse code z^m is obtained similar as Eq.(12). Therefore, the time complexity is $O(p^m K)$.

Binary Embedding: The linear transformation from a sparse code z^m to a binary hash code is achieved by the hash functions in Eq.(13) which takes $O(KL)$ time.

An intuitive comparison of DCDH and other state-of-the-art cross-media hashing methods on off-line training and on-line testing time are demonstrated in Figure 2. We can see that our DCDH and MV-DCDH require the least time in the training stage and is also very efficient in the testing stage.

5. EXPERIMENTS

In our experiments, we evaluate the performance of our DCDH. We first introduce the data set, evaluation criteria and the parameter setting we used in the experiments. Then, we compare our DCDH with other *state-of-the-art* methods and analyze the results. Finally, we further investigate the learned coupled dictionary space to explain why our DCDH and MV-DCDH achieve the superior performance.

5.1 Experimental Setup

We use two real-world data sets "Wikipedia-Picture of the Day"(abbreviated as Wiki-Potd) [2] and NUS-WIDE[3]. Both data sets are bi-modal containing images and texts.

The Wiki-Potd data set consists of 2866 Wikipedia documents. Each document contains one text-image pair. All documents are labeled by one of 10 semantic categories. For the image modality, we extract 1000-D Bag of visual words (BoVW) and 512-D GIST descriptors for each image. For the text modality, we calculate the frequency of all words in the data set and select the most representative words to quantize all texts into 5,000-D Bag-of-Words (BoW).

The NUS-WIDE data set contains 269,648 labeled images and is manually annotated with 81 categories. Each image with its annotated tags in NUS-WIDE can be taken as a pair of image-text data. To guarantee that each category has abundant training samples, we select those pairs that belong to one of the 10 largest categories (e.g., 'sky', 'buildings', 'person') with each pair exclusively belonging to one of the 10 categories (discrimination on concepts are required when learning coupled dictionaries.). For the image modality, over three types of visual features are extracted for each image (i.e., 500-D BoVW, 255-D Color Moments, 128-D Wavelet Texture). For the text modality, the corresponding labels of each image are represented by a 1,000-D BoW.

[2] `http://www.svcl.ucsd.edu/projects/crossmodal/`
[3] `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm`

Table 2: The details of the data sets used in the experiments

Data Set	NUS-WIDE	Wiki-Potd
Image modality	BoVW(500-D) CM(255-D) Wavelet(128-D)	BoVW(1000-D) GIST(512-D) -
Text modality	BoW(1000-D)	BoW(5000-D)
Data set size	60641	2866
Training set size	3000	1000
Validation set size*	2000 /10000	866/866
Testing set size*	2000 / 47641	866/1000

* Partitions are ordered by query/database set respectively. The query set are random sampled from the database set.

The details of the two data sets are shown in Table 2.

To evaluate the performance of the cross-media retrieval results, we adopt the mean average precision (MAP) and mean average top-R precision (MAP@R) defined in [13].

5.2 Compared Methods

We perform three types of retrieval schemes in the experiments : 1) Image-query-Texts: use image queries to retrieve relevant texts. 2) Text-query-Images: use text queries to retrieve relevant images. 3) Image-query-Images: use image queries to retrieve relevant images. For the first two retrieval schemes, we compare with the state-of-the-art cross-media hashing methods CMSSH [3], CVH [9], MLBE [28], LCMH [30]; for the third retrieval scheme, we additionally compare with some uni-modal hashing approaches: SH [23], KLSH [8], AGH [11] and a multi-view hashing approach MFH [17]. The reason why we don't compare DCDH with the approaches in [31] and [25] is that they can not generate *compact* and *binary* hash codes, thus their performance can not be fairly evaluated under the same settings.

Our DCDH method and its multi-view enhancement are denoted as DCDH and MV-DCDH, respectively.

For the Image-query-Texts and Text-query-Images retrieval schemes, the performances of CMSSH, CVH, MLBE, LCMH, DCDH, MV-DCDH are compared. Except for our MV-DCDH which induces multi-view features, the remaining methods take the BoVW descriptors for the image modality and BoW for text modality; for MV-DCDH, the multi-view features are utilized for image modality.

For the Image-query-Images retrieval scheme, we compare with all the counterparts aforementioned. It is notable that for MFH, the multi-view features are exploited.

5.3 Parameters Sensitivity

There are four parameters in DCDH: the k-NN of the intra-modality; λ', γ' in Eq.(11) when learning the coupled dictionaries; the size of the coupled dictionaries K.

Following the prior settings in [7, 10], λ' and γ' are set to 10 and 1 respectively throughout the experiments.

We fix the code length $L = 24$ and evaluate the average MAP variations (the average MAP scores of Image-query-Texts and Text-query-Images) in terms of K and k-NN on the validation set. The tested combinations are $K = \{50, 100, 200, 300, 400, 500\}$ and k-NN $= \{5, 10, 20, 30, 50\}$, The optimal combination on NUS-WIDE is k-NN $= 5$, $K =$

Table 3: The MAP performance comparison on the NUS-WIDE data set with code length L varying from 8 to 40. The items in bold are the two best results, and the results with asterisk are the best.

Task	Methods	Hash code length		
		$L = 8$	$L = 24$	$L = 40$
Image query Texts	CVH	0.3144	0.3139	0.3140
	CMSSH	0.3233	0.3131	0.3140
	MLBE	0.3142	0.3138	0.3119
	LCMH	0.3163	0.3117	0.3144
	DCDH	**0.3608**	**0.3573**	**0.3568**
	MV-DCDH	**0.3645***	**0.3627***	**0.3608***
Task	**Methods**	**Hash code length**		
		$L = 8$	$L = 24$	$L = 40$
Text query Images	CVH	0.3158	0.3152	0.3144
	CMSSH	0.3373	0.3309	0.3287
	MLBE	0.3133	0.3156	0.3167
	LCMH	0.3142	0.3124	0.3133
	DCDH	**0.3452**	**0.3559**	**0.3554**
	MV-DCDH	**0.3640***	**0.3629***	**0.3604***

Table 4: The MAP performance comparison on the Wiki-Potd data set.

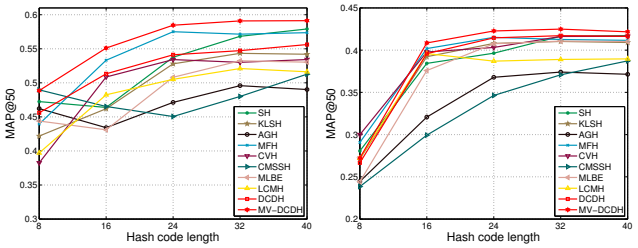
Task	Methods	Hash code length		
		$L = 8$	$L = 24$	$L = 40$
Image query Texts	CVH	0.1490	0.1407	0.1381
	CMSSH	0.1448	0.1407	0.1431
	MLBE	0.1445	0.1359	0.1371
	LCMH	0.1267	0.1273	0.1258
	DCDH	**0.1821**	**0.1985**	**0.1934**
	MV-DCDH	**0.2017***	**0.2010***	**0.1997***
Task	**Methods**	**Hash code length**		
		$L = 8$	$L = 24$	$L = 40$
Text query Images	CVH	0.1435	0.1361	0.1351
	CMSSH	0.1412	0.1364	0.1380
	MLBE	0.1449	0.1344	0.1363
	LCMH	0.1260	0.1237	0.1235
	DCDH	**0.1606**	**0.1648**	**0.1620**
	MV-DCDH	**0.1745***	**0.1734***	**0.1716***

100 and k-NN $= 20$, $K = 300$ on Wiki-Potd. These settings are adopted in the following experiments.

5.4 Performance Comparisons

We compare our DCDH and its extension MV-DCDH with the three following methods: CMSSH [3], CVH [9] and MLBE [28]. We evaluate the cross-media retrieval performance with code length varying from 8 to 40 and report results in terms of MAP in Table 3 and 4, respectively. Moreover, we report the Image-query-Images retrieval performance in terms of MAP@50 in Figure 3. The reason why we choose MAP@50 rather than MAP for the Image-query-Images task is that the differences of MAP scores, especially for the counterparts, are not statically significant, which makes it difficult to illustrate them explicitly in a line graph.

The MAP scores on Wiki-Potd data set is generally lower than that on NUS-WIDE even if the Wiki-Potd data set contains rich textual information. This may be explained as that the categories of Wiki-Potd data set (e.g., 'art', 'biology') is too general, so the feature vector can not precisely capture its corresponding semantic meaning.



(a) NUS-WIDE (b) Wiki-Potd

Figure 3: The performance comparison of Image-query-Images on two data sets.

It can be noted that DCDH significantly outperforms the counterparts over different code lengths: 2% $\sim$ 7% on Wiki-Potd data set and 2% $\sim$ 5% on NUS-WIDE, respectively. The improvement is due to the effectiveness of the sparse representation over the discriminative coupled dictionaries. All the counterparts simply project the data from different modalities into a shared Hamming space using the learned hash functions. However, the meaning of the shared Hamming space is ambiguous and cannot well clarify the semantic information of the data. By contrast, our DCDH exploits the category information when learning the coupled dictionaries, thus making the coupled dictionary space semantic interpretable. By utilizing the distribution of the sparse codes, the manifold structure of the dictionary space is also preserved in the embedding Hamming space. Therefore, the binary hash codes represent the semantic information of the data and lead to superior cross-media retrieval performance. Moreover, we find that the incorporation of the multi-view features on the image modality, i.e., MV-DCDH, produces a significant improvement over DCDH on both of the Image-query-Texts and the Text-query-Images tasks. This observation is explained as that exploiting multiple features over the "weak" modality (image modality) gives better understanding of the semantics of the images. This observation also verifies our hypothesis that a balanced representation is important in cross-media retrieval.

Although our main goal is cross-media retrieval, we can readily use the hash functions we learned to perform uni-modal retrieval. That is to say, all the cross-media hashing approaches can be adapted to uni-modal hashing. We conduct the task of Image-query-Images and report the performance of the aforementioned cross-media hashing approaches. Besides, we add some state-of-the-art uni-modal hashing and multi-view hashing approaches to perform fair experimental comparison. The results are shown in Figure 3.

From Figure 3, we find that all the cross-media hashing approaches except CMSSH achieve reasonable performance. This is due to the fact that the learned hash functions for image modality can borrow strength from text modality. The observation for the poor performance of CMSSH in this task may be explained as the lack of intra-modality preservation when learning hash functions in CMSSH.

In addition, since MV-DCDH induces multi-features, we also add MFH [17] into comparison which also exploits multi-features when learning hash functions. The results show that MFH outperforms the other uni-modal hashing approaches and most of the cross-media hashing approaches, which demonstrates the effectiveness of multi-views in image understanding. Our MV-DCDH slightly outperforms

Table 5: Topic words and corresponding category labels of some selected dictionary atoms on the Wiki-Potd data set.

Categories	Topic Words
Biology	Skull Dinosaurs Bone Fossils Prey
	Whales Killer Meat Oil Atlantic
Sport	Goals Hockey Players Montreal NHL
	Football Yard Bowl Champion Quarter
Warfare	Natives Crops Australian Infantry Landing
	Soviet, Moscow Marshall Defense Battle
Media	Theater Broadway Movie Actor Disc
Geography	Creek Tree Parks Ridge Forest

the MFH and achieves the overall best performance in the Image-query-Images task.

5.5 The Discriminative Capability of the Coupled Dictionary Space

The results over both the data sets above outline the superior performance of DCDH over the other cross-media hashing approaches. This superiority mainly owes to the aptitude of the discriminative capability of the coupled dictionary space in DCDH. To verify our hypothesis, we investigate the coupled dictionary space. We choose the Wiki-Potd data set since it has rich textual information and is convenient for illustrations. To show the effect of discriminative capability of the coupled dictionary space, we design a non-discriminative version of DCDH by simply setting $\lambda' = 0$. The rest settings are same as the ones used aforementioned ($K = 300$ and k-NN $= 20$).

Denote the learned coupled dictionaries for image and text modalities as D^x amd D^y, respectively. Z^x and Z^y are the sparse codes of the test data set in two modalities by the corresponding dictionaries. To measure the discrimination of the sparse codes, we define a metric called *Discriminative Degree* (abbreviated as DD) as follows:

$$DD(Z) = \frac{1}{N}\sum_{i=1}^{N} P(z_i) \tag{15}$$

where N is the number of the testing samples. $P(z_i) = 1$ if at least one of the selected dictionary atoms of the sparse code z_i indicates the true category label of the i-th data point and 0 otherwise.

Moreover, DD can also be used to measure the coupling degree of the sparse codes from different modalities. Denote the dot product of Z^x and Z^y as Z^{xy} and $DD(Z^{xy})$ reflects the one-to-one correspondence of the paired dictionary atoms .

Figure 4 shows the comparison results. The "Random" method indicates that the dictionaries are randomly generated; The "Non-discriminative" method is the aforementioned DCDH with $\lambda' = 0$. The "Discriminative" and "MV-Discriminative" methods correspond to our DCDH and MV-DCDH. From the results, we get four observations: 1) the sparse codes of the text modality is more semantically discriminative than the ones from the image modality (even when we do not impose the discriminative constraints); 2) The introduction of side information improves the discriminative capability especially for the image modality (without the category side information, the DD score of the image modality is almost equal to the random method); 3) The coupling degree is significantly improved when imposing the discriminative constraint. This is mainly due to the better representation of the image modality; 4) Exploiting multi-view features further improves the performance on both the understanding of the image modality and the coupling degree.

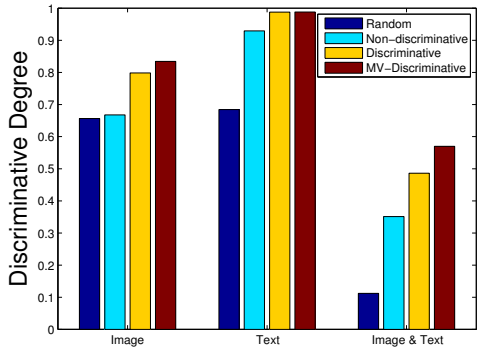


Figure 4: The discriminative capability comparison

To give an intuitive illustration of the learned dictionary, we give an insight into the textual dictionary D^y (the image dictionary D^x is learned from the BoVW, which is difficult to illustrate). Since each dictionary atom d_k^y is obtained by clustering a portion of BoW features, the values of the elements in d_k^y measure the occurrence frequency of the words. Naturally, we can use the elements with largest values in d_k^y to represent the topic words of this dictionary atom d_k^y. Moreover, each d_k^y is learned with a discriminative constraint and has a dominated category label, we demonstrate the relation between the category label and topic words of the dictionary atoms in Table 5 (topic words correspond to the 5 largest elements of the selected dictionary atoms). From the results, we can find that the topic words for each dictionary atom indeed reflect some certain semantic information and are consistent with its belonging category. Besides, two dictionary atoms belong to the same category have an explicit disparity in semantics. e.g., the two atoms from the category "Biology" describe different topics of "Biology": the first topic is about dinosaurs and the second one is about whale killing. This observation can be explained as the collaborative effect of the compact function and discriminative function. The discriminative function encourages the topics to be classified into the correct categories and the compact function encourages each topic to reflect an individual aspect of its corresponding category.

6. CONCLUSIONS

In this paper, we propose a discriminative coupled dictionary hashing (DCDH) approach for fast cross-media retrieval. Our DCDH is two-stage in that we first learn coupled dictionary for each modality discriminatively with the side information of category labels, so that the data from different modalities are represented as the sparse codes in a shared semantically discriminative dictionary space. Afterwards, the sparse codes are mapped to binary hash codes by the learned unified hash functions to support fast cross-media retrieval. Extensive experiments on two real-world data sets demonstrate the superior performance of DCDH over the existing state-of-the-art hashing approaches.

Moreover, we conjecture that a balanced cross-media representation benefits the cross-media retrieval performance.

Therefore, we extend DCDH to MV-DCDH which introduces multi-view features on the relatively "weak" modalities (i.e., the image modality in our experiments) to obtain a balanced representation. The experimental results verify the effectiveness of MV-DCDH.

7. ACKNOWLEDGEMENT

This work is supported in part by National Basic Research Program of China (2012CB316400), NSFC (No.61128007), 863 program (2012AA012505), the Fundamental Research Funds for the Central Universities and Chinese Knowledge Center of Engineering Science and Technology (CKCEST) and Program for New Century Excellent Talents in University. Dr.Qi Tian is also supported by ARO grant W911NF-12-1-0057, Faculty Research Award by NEC Laboratories of America, and 2012 UTSA START-R Research Award respectively. Dr. Jiebo is also supported by Google Faculty Research Awards.

8. REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionries for sparse representation. *IEEE Trans.Signal Processing*, 54(11):4311–4322, 2006.

[2] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, pages 459–468, 2006.

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

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Sstatistics*, 32(2):407–499, 2004.

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

[6] K. Jia, X. Tang, and X. Wang. Image transformation based on learning dictionaries across image spaces. *IEEE Trans.Pattern Anal. Mach. Intell.*, 2012.

[7] Z. Jiang, G. Zhang, and L. S. Davis. Submodular dictionary learning for sparse coding. In *CVPR*, pages 3418–3425, 2012.

[8] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, pages 2130–2137, 2009.

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

[10] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In *CVPR*, pages 2097–2104, 2011.

[11] W. Liu, J. Wang, S. Kumar, and S. Chang. Hashing with graphs. In *ICML*, pages 1–8, 2011.

[12] Y. Liu, F. Wu, Y. Yi, Y. Zhuang, and A. Hauptman. Spline regression hashing for fast image search. *IEEE Trans. Image Processing*, 2012.

[13] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang. A low rank structural large margin method for cross-modal ranking. In *SIGIR*, pages 433–442. ACM, 2013.

[14] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functionsąli. *Mathematical Programming*, 14(1):265–294, 1978.

[15] M. Ou, P. Cui, F. Wang, J. Wang, W. Zhu, and S. Yang. Comparing apples to oranges: a scalable solution with heterogeneous hashing. In *SIGKDD*, pages 230–238, 2013.

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

[17] J. Song, Y. Yang, Z. Huang, H. Shen, and R. Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM MM*, pages 423–432, 2011.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[19] C. Wang and S. Mahadevan. A general framework for manifold alignment. In *AAAI*, 2009.

[20] J. Wang, S. Kumar, and S. Chang. Semi-supervised hashing for scalable image retrieval. In *CVPR*, pages 3424–3431, 2010.

[21] Q. Wang, D. Zhang, and L. Si. Semantic hashing using tags and topic modeling. In *SIGIR*, pages 213–222, 2013.

[22] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *CVPR*, pages 2216–2223, 2012.

[23] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

[24] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans.Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[25] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang. Sparse multi modal hashing. *IEEE Trans. Multimedia*, 16(2):427–439.

[26] D. Zhang, F. Wang, and L. Si. Composite hashing with multiple information sources. In *SIGIR*, pages 225–234, 2011.

[27] D. Zhang, J. Wang, D. Cai, and J. Lu. Self-taught hashing for fast similarity search. In *SIGIR*, pages 18–25, 2010.

[28] Y. Zhen and D. Yeung. A probabilistic model for multimodal hash function learning. In *SIGKDD*, 2012.

[29] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In *NIPS*, pages 1385–1393, 2012.

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

[31] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013.

[32] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Trans. Multimedia*, 10(2):221–229, 2008.