# Supervised Coupled Dictionary Learning with Group Structures for Multi-Modal Retrieval

**Yueting Zhuang** and **Yanfei Wang** and **Fei Wu** and **Yin Zhang** and **Weiming Lu**
College of Computer Science
Zhejiang University, China
{yzhuang,yanfeiwang07,wufei,zhangyin98,luwm}@zju.edu.cn

## Abstract

A better similarity mapping function across heterogeneous high-dimensional features is very desirable for many applications involving multi-modal data. In this paper, we introduce coupled dictionary learning (DL) into supervised sparse coding for multi-modal (*cross-media*) retrieval. We call this *S*upervised coupled-dictionary *l*earn*i*ng with group structures for *M*ulti-*M*odal retrieval (SliM$^2$). SliM$^2$ formulates the multi-modal mapping as a constrained dictionary learning problem. By utilizing the intrinsic power of DL to deal with the heterogeneous features, SliM$^2$ extends *uni-modal* DL to *multi-modal* DL. Moreover, the label information is employed in SliM$^2$ to discover the shared structure inside intra-modality within the same class by a mixed norm (i.e., $\ell_1/\ell_2$-norm). As a result, the multi-modal retrieval is conducted via a set of jointly learned mapping functions across multi-modal data. The experimental results show the effectiveness of our proposed model when applied to cross-media retrieval.

## Introduction

Similarity search, a.k.a. nearest neighbor search, is a fundamental problem and has enjoyed success in many applications of data mining, database, and information retrieval. Nevertheless, most of the similarity search algorithms are only conducted in the *uni-modal* data setting, which are restricted to retrieve the similar data with the same modality as query data. Nowadays, many real-world applications involve multi-modal data, where information inherently consists of data with different modalities, such as a web image with loosely related narrative text descriptions, or a news article with paired text and images. Therefore, it is desirable to support similarity search for multi-modal data (*i.e., cross-media retrieval*), e.g., the retrieval of textual documents in response to a query image or vice versa (Wu, Zhang, and Zhuang 2006) (Zhuang, Yang, and Wu 2008). Multi-modal retrieval is very important to many applications of practical interest, such as finding some detailed textual documents of a tourist spot that best match a given image, obtaining a set of images that best visually illustrate a given text, or searching similar results by a set of combined texts and images.



To the best of our knowledge, there are generally two kinds of approaches to boost cross-modal retrieval: one is canonical correlation analysis (CCA) (Hotelling 1936) and its variants. For examples, after the maximally correlated subspace of text and image features is obtained by CCA, logistic regression is employed to cross-media retrieval in (Rasiwasia et al. 2010). A supervised extension of CCA, referred as generalized multiview analysis (GMA), was proposed in (Sharma et al. 2012) for cross-media retrieval. These existing CCA-based approaches attempt to enforce a strong assumption among the multi-modal data, i.e., the different modalities have a common or a shared subspace. However, this assumption is too restricted to some extent for analysis of multi-modal data in real-world setting. For example, given a pair of image and text, the image probably contains a considerable amount of information not related to its corresponding text, and it is not even guaranteed that the text is related at all to the visual content of the image.

Another kind of approaches for multi-modal retrieval are extensions of Latent Dirichlet Allocation (LDA). Following the seminal work of Blei et al.(Blei, Ng, and Jordan 2003), Latent Dirichlet Allocation (LDA) has been extended to learn the joint distribution of multi-modal data (e.g., texts and images) such as Correspondence LDA (Corr-LDA) (Blei and Jordan 2003), Topic-regression Multi-modal LDA (tr-mmLDA) (Putthividhy, Attias, and Nagarajan 2010), Multi-field Correlated Topic Modeling (mf-CTM) (Salomatin, Yang, and Lad 2009) and Hierarchical Dirichlet Process(HDP) based LDA (Virtanen et al. 2012). These aforementioned approaches tend to model the correlations of multi-modal data at latent semantic (*topic*) level across modalities. Therefore, they either assume that all modalities share same topic proportions, or have one-to-one topic correspondences, or have commonly shared topics. Nevertheless, those assumptions inherently restrain a more flexible application of cross-media retrieval in the setting involved uncontrolled multi-modal data .

On the other hand, when the class labels (categories) of multi-modal data are available, it is natural to assume that intra-modality data within the same class (category) shares some common *aspects*. For examples, images from the "architecture" category have similar low-level visual features (such as geometric regularities and patches of uniform color (Todorovic and Nechyba 2004) ), and textual documents from "biology" have overlapping words (e.g., cells and genetics). Therefore, it is appropriate to utilize the class labels to learn the discriminately shared components for intra-modal data from the same category. Motivated by the fact that dictionary learning (DL) methods have the intrinsic power of dealing with the heterogeneous features by generating different dictionaries for multi-modal data, this paper tends to study on jointly learning multi-modal dictionaries in a supervised setting, and simultaneously mining the shared structures inside each intra-modality from the same classes.

There are some existing DL approaches for multi-modal data. Method was proposed in (Monaci et al. 2007) to learn multi-modal dictionaries for audiovisual data. This model, however, can only deal with synchronous temporal signals. A dictionary learning approach is proposed in (Jia, Salzmann, and Darrell 2010) to factorize the latent space across modalities into shared components (to all modalities) and private parts (to each modality). The assumption in (Jia, Salzmann, and Darrell 2010) that assumes a unique sparse coefficient across all the modalities is still too restricted to multi-modal data in real-world applications.

Inspired by the recently proposed idea of (semi-)coupled dictionary learning (CDL) for image super-resolution (Jia,Tang, and Wang 2012) and photo-sketch synthesis (Wang et al. 2012), which suggest that one pair of image patches from different domains (low resolution *vs* high resolution, or photo *vs* sketch) has the same dictionary entries or has a mapping function between the reconstructed sparse coefficients, this paper proposes *S*upervised coupled dictionary *l*earn*i*ng with group structures for *M*ulti-*M*odal retrieval (SliM$^2$). SliM$^2$ extends uni-modal DL to multi-modal DL and jointly learns a set of mapping functions across different modalities. Furthermore, SliM$^2$ utilizes the label information to discover the shared structures inside intra-modalities from the same classes.

## The Model of SliM$^2$

In this section, we first briefly review sparse coding and its extensions, then we present the formulation of SliM$^2$. At last, SliM$^2$ is conducted for multi-modal retrieval.

### Dictionary Learning and Its Extensions

The modeling of data with the linear combinations of a few elements from a learned dictionary has been the focus of much recent research (Olshausen, Field, and others 1997) (Wright et al. 2009). The essential challenge to be resolved in sparse coding is to develop an efficient approach with which each sample can be reconstructed from a 'best dictionary' with a 'sparse coefficients'.

Let $\mathbf{X} \in R^{p\times n}$ be the data matrix to be reconstructed, $\mathbf{D} \in R^{p\times k}$ the learned dictionary and $\alpha \in R^{k\times n}$ the sparse reconstruction coefficients (also known as *sparse codes*), where $p$, $n$ and $k$ are the dimensions of feature space, the number of data samples and the size of the dictionary respectively. The formulation of sparse coding can be expressed as follows:

$$\begin{aligned} &\min_{\mathbf{D},\alpha} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\alpha\|_F^2 + \lambda\Psi(\alpha) \\ &s.t. \quad \|\mathbf{d}_i\| \le 1,\ \forall i, \end{aligned} \tag{1}$$

where $\Psi(\alpha)$ represents the imposed penalty over sparse codes $\alpha$ and $\mathbf{d}_i$ is one of the dictionary atoms of $\mathbf{D}$. Typically, the $l_1$ norm is conducted as a penalty to explicitly enforce sparsity on each sparse codes $\alpha_j$ ($\alpha_j \in \alpha(j = 1,\ldots,N)$) (Tibshirani 1996) (Jia, Salzmann, and Darrell 2010) as follows

$$\Psi(\alpha) = \sum_{j=1}^{N} \|\alpha_j\|_1 \ . \tag{2}$$

The above classical data-driven approach to dictionary learning is well adapted to reconstruction tasks such as restoring a noisy signal. In order to learn a discriminative sparse model instead of purely reconstructive one, sparse coding is extended into supervised sparse coding (Mairal et al. 2008). In real-word setting, different data can be naturally designated into different groups, a mixed-norm regularization ($\ell_1/\ell_2$-norm) can be conducted in sparse coding to achieve sparsity as well as to encourage the reconstruction of samples from the same group by the same dictionary atoms, which is named as *group sparse coding* in (Bengio et al. 2009).

If all of the images in one class (category) is taken as a group, as stated before, it is appropriate to assume that when a set of dictionary atoms has been selected to represent one image of a given category, the same dictionary atoms could be used to represent other images of the same category (Bengio et al. 2009). The formulation of group sparse coding is as follows:

$$\begin{aligned} &\min_{\mathbf{D},\alpha} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\alpha\|_F^2 + \lambda\sum_{l=1}^{J}\sum_{i=1}^{k}\|\alpha_{i,\Omega_l}\|_2 \\ &s.t. \quad \|\mathbf{d}_i\| \le 1,\ \forall i, \end{aligned} \tag{3}$$

where $J$ is the number of classes (groups), $\Omega_l$ represents the indices of the examples that belong to the $l$-th class ($l$-th group), and $\alpha_{:,\Omega_l}$ is the coefficient matrix associated to examples in the $l$-th group.

### The Formulation of SliM$^2$

Suppose that we have a labeled training set of $N$ pairs of correspondence data with $M$ modalities from $J$ classes: $\{(x_i^{(1)},\cdots,x_i^{(M)},l_i) : i = 1,\ldots,N\} \in \{(\mathbf{X}^{(1)},\cdots,\mathbf{X}^{(M)},\mathbf{L})\}$. $\mathbf{X}^{(m)} \in R^{P_m\times N}$ ($1 \le m \le M$) is $P^m$-dimensional data from the $m$-th modality, $l_i = (l_{i1},\ldots,l_{iJ})' \in \{0,1\}^J$ is the corresponding class label, $l_{ij} = 1$ if the $i$-th data $x_i = (x_i^{(1)},\cdots,x_i^{(M)})$ belongs to the $j$th class and $l_{ij} = 0$ otherwise. Here, the $i$-th data $x_i$ only belongs to a single class: $\sum_{j=1}^{J} l_{ij} = 1$.

We have seen from Eq.(3) that group sparse coding is a way for uni-modal dictionary learning when the input signals are naturally assigned into different groups. Of particular interest to us in this paper is modeling the relationships between *multi-modal* data rather than the independent dictionary learning from *uni-modal* data. In order to resolve this issue, we resort to semi-coupled DL (Wang et al. 2012) for a mapping between reconstruction coefficients. The underlying motivation behind our SliM$^2$ has two points: a) jointly learn dictionaries for each modality data and a relatively simple mapping function across modalities; b) discover the shared structures for each intra-modality data from the same class *via* a mixed norm (i.e., $\ell_1/\ell_2$-norm).

SliM$^2$ aims to jointly learn a set of dictionaries for $M$ modality data respectively, i.e., $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ with $\mathbf{D}^{(m)} \in R^{P_m \times K}$ and their corresponding reconstruction coefficients $A = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(M)}\}$ with $\mathbf{A}^{(m)} \in R^{K \times N}$, where $K$ is the size of the dictionaries (the number of *atoms* in dictionary). In order to conduct the multi-modal retrieval, we assume there exists a set of linear mappings $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ with $\mathbf{W}^{(m)} \in R^{K \times K}$ between sparse codes. The objective function of our proposed SliM$^2$ is formulated as follows:

$$\begin{aligned}
&\min \sum_{m=1}^{M} \|\mathbf{X}^{(m)} - \mathbf{D}^{(m)}\mathbf{A}^{(m)}\|_F^2 + \sum_{m=1}^{M}\sum_{l=1}^{J} \lambda_m \|\mathbf{A}^{(m)}_{:,\Omega_l}\|_{1,2} \\
&+ \beta \sum_{m=1}^{M}\sum_{n\neq m} \|\mathbf{A}^{(n)} - \mathbf{W}^{(m)}\mathbf{A}^{(m)}\|_F^2 + \gamma \sum_{m=1}^{M} \|\mathbf{W}^{(m)}\|_F^2 \\
&s.t. \quad \|\mathbf{d}_k^{(m)}\| \leq 1, \quad \forall k, \ \forall m,
\end{aligned} \tag{4}$$

where $\mathbf{A}_{:,\Omega_l}$ is the coefficient matrix associated to those intra-modality data belonging to the $l$-th class. For an arbitrary matrix $\mathbf{A} \in \mathbf{R}^{\mathbf{k}\times\mathbf{n}}$, its $\ell_1/\ell_2$-norm is defined as

$$\|\mathbf{A}\|_{1,2} = \sum_{i=1}^{k} \sqrt{\sum_{j=1}^{n} \mathbf{A}_{ij}^2}\ . \tag{5}$$

Here, $\beta$,$\gamma$ and $\lambda_m (m = 1, \ldots, M)$ are tuning parameters denoting the weights of each term in Eq.(4). It is obvious that data in the $m$-th modality space can be mapped into the $n$-th modality space by the learned $W^{(m)}$ according to $\|\mathbf{A}^{(n)} - \mathbf{W}^{(m)}\mathbf{A}^{(m)}\|_F^2$, therefore, the computation of multi-modal similarity is achieve in SliM$^2$.

The degree of sparsity for data across modalities could be different due to their heterogeneity with high-dimensional settings. As a result, different $\lambda_m (m \in \{1, \ldots, M\})$ is employed in Eq.(4) to control the degree of sparsity of the sparse codes respectively for $M$ modality data.

It can be observed from Eq.(4) that the proposed SliM$^2$ not only jointly minimizes the reconstruction error of data across modalities, but also independently encourages to utilize same dictionary *atoms* for the reconstruction of the intra-modality data from the same class.

## The Optimization of SliM$^2$

The aforementioned objective function in Eq.(4) is non-convex and non-smooth, but it is convex to each set of $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$, $A = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(M)}\}$ and $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ when the other two are fixed. Therefore, in practice, we can develop an iterative algorithm to optimize the variables alternatively. This approach is called the alternative minimization and is widely used in many applications such as (Kang, Grauman, and Sha 2011) and (Jia, Tang, and Wang 2012).

First, we fix $D$ and $W$ to optimize $A$. We initialize $W$ as identity matrix and $D$ using the dictionary learning algorithm in (Mairal et al. 2010) respectively. With $D$ and $W$ fixed, the optimization of $A$ can be obtained as follows:

$$\begin{aligned}
&\min_A \sum_{m=1}^{M} \|\mathbf{X}^{(m)} - \mathbf{D}^{(m)}\mathbf{A}^{(m)}\|_F^2 + \sum_{m=1}^{M}\sum_{l=1}^{J} \lambda_m \|\mathbf{A}^{(m)}_{:,\Omega_l}\|_{1,2} \\
&+ \beta \sum_{m=1}^{M}\sum_{n\neq m} \|\mathbf{A}^{(n)} - \mathbf{W}^{(m)}\mathbf{A}^{(m)}\|_F^2\ .
\end{aligned} \tag{6}$$

Eq.(6) is a problem of multi-modal group sparse coding and we use block-coordinate descent (Qin, Scheinberg, and Goldfarb 2010) (Friedman, Hastie, and Tibshirani 2010) to solve it.

After obtaining $A$, we then update the dictionaries $D$ as follows:

$$\begin{aligned}
&\min_D \sum_{m=1}^{M} \|\mathbf{X}^{(m)} - \mathbf{D}^{(m)}\mathbf{A}^{(m)}\|_F^2 \\
&s.t. \quad \|\mathbf{d}_k^{(m)}\| \leq 1, \quad \forall k, \ \forall m,
\end{aligned} \tag{7}$$

This is a quadratically constrained quadratic program (QCQP) problem which can be solved using the method presented in (Yang et al. 2010).

Finally, we update $W$ as follows:

$$\begin{aligned}
&\min_W \sum_{m=1}^{M}\sum_{n\neq m} \|\mathbf{A}^{(n)} - \mathbf{W}^{(m)}\mathbf{A}^{(m)}\|_F^2 \\
&+ (\gamma/\beta) \sum_{m=1}^{M} \|\mathbf{W}^{(m)}\|_F^2,
\end{aligned} \tag{8}$$

This is a set of ridge regression problem and can be worked out as follows:

$$\mathbf{W}^{(m)} = \mathbf{A}^{(n)}{\mathbf{A}^{(m)}}^T(\mathbf{A}^{(m)}{\mathbf{A}^{(m)}}^T + (\gamma/\beta)\cdot\mathbf{I})^{-1}, \tag{9}$$

where $\mathbf{I}$ is the identity matrix. The above procedure iterates until the convergences of $A$, $D$ and $W$ are achieved.

## SliM$^2$ for multi-modal retrieval

Given a query $x_q^{(m)} \in R^{P_m}$ from $m$-th modality , suppose we are looking for its similar data from the $n$-th modality.

Now we have jointly learned the dictionary for each modality data $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ and a set of mapping functions $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$. For the query data $\mathbf{x}_q^{(m)}$, we need to map $\mathbf{x}_q^{(m)}$ into the space of $n$-th modality data. With the initialization as follows:

$$\begin{aligned}
\alpha_q^{(m)} &= \min_{\alpha_q} \frac{1}{2}\|\mathbf{x}_q^{(m)} - \mathbf{D}^{(m)}\alpha_q^{(m)}\|_F^2 + \lambda\|\alpha_q^{(m)}\|_1 \\
\alpha_r^{(n)} &= \mathbf{W}^{(m)}\alpha_q^{(m)} \\
\mathbf{x}_r^{(n)} &= \mathbf{D}^{(n)}\alpha_r^{(n)},
\end{aligned} \tag{10}$$

**Algorithm 1** **The optimization of SliM$^2$**

**Input** The labeled training set of $N$ pairs data with $M$ modalities from $J$ classes $\{(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(M)}, l_i)\} \in \{(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(M)}, \mathbf{L})\}$.

1: Initialize $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ and $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$,

2: Optimize $A = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \cdots, \mathbf{A}^{(M)}\}$ by Eq.(6),

3: Update $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ with other variables fixed using Eq.(7),

4: Update $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ with other variables fixed using Eq.(9),

5: Repeat 2-4 until convergence.

**Output** multi-modal dictionaries $D$ and a set of mapping functions $W$

we then obtain optimized $\hat{\alpha}_r^{(n)}$ and $\hat{\alpha}_q^{(m)}$ as follows:

$$\min_{\hat{\alpha}_q^{(m)}, \hat{\alpha}_r^{(n)}} \|\mathbf{x}_q^{(m)} - \mathbf{D}^{(m)}\alpha_q^{(m)}\|_F^2 + \|\mathbf{x}_r^{(n)} - \mathbf{D}^{(n)}\alpha_r^{(n)}\|_F^2 + \beta\|\alpha_r^{(n)} - \mathbf{W}^{(m)}\alpha_q^{(m)}\|_F^2 + \lambda_m\|\alpha_q^{(m)}\|_1 + \lambda_n\|\alpha_r^{(n)}\|_1 \,. \tag{11}$$

The query data $\mathbf{x}_q^{(m)}$ can be mapped into $n$-th modality data $\hat{\mathbf{x}}_r^{(n)}$ as follows:

$$\hat{\mathbf{x}}_r^{(n)} = \mathbf{D}^{(n)}\hat{\alpha}_r^{(n)} \,. \tag{12}$$

Thus, all of data in the $n$-th modality which has the least distances to $\mathbf{x}_r^{(n)}$ is ranked as the retrieved results of the query data.

We summarize the optimization of SliM$^2$ in Algorithm 1 and multi-modal retrieval by the SliM$^2$ in Algorithm 2.

**Algorithm 2** **The multi-modal retrieval by SliM$^2$**

**Input** The learned multi-modal dictionaries $D = \{\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \cdots, \mathbf{D}^{(M)}\}$ and a set of mapping functions $W = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \cdots, \mathbf{W}^{(M)}\}$ from training data and query data $\mathbf{x}_q^{(m)} \in R^{P_m}$ in the $m$-th modality

1: Initialize $\alpha_q^{(m)}$,$\alpha_r^{(n)}$ and corresponding retrieval $\mathbf{x}_r^{(n)}$ using Eq.(10),

2: Optimize $\hat{\alpha}_q^{(m)}$,$\hat{\alpha}_r^{(n)}$ with other variables fixed using Eq.(11),

3: Update $\hat{\mathbf{x}}_r^{(n)}$ using Eq.(12),

4: Repeat 2-3 until convergence.

5: the ranked neighbors of $\hat{\mathbf{x}}_r^{(n)}$.

**Output** The retrieved similar data in the $n$-th modality

# Experiments

In this section, we evaluate the performance of our proposed SliM$^2$ when applied to cross-media retrieval. We first introduce the data sets and evaluation criterions we adopted, then we elaborate parameter setting and tuning in our experiments. At last, we compare SliM$^2$ with other state-of-the-art algorithms and demonstrate the results.

## Data Sets

One of our experimental data sets is the Wiki Text-Image data (Rasiwasia et al. 2010). Wiki Text-Image contains 2173/693(training/testing) text-image pairs from ten different categories. After SIFT features (Lowe 1999) are extracted, k-means clustering is conducted to obtain the representation of bag-of-visual-words (abbreviated as *BoVW*) (Fei-Fei, Fergus, and Perona 2004) for each image. The term frequency is used to obtain the representation of bag-of-textual-words (abbreviated as *BoW*) for each text. Since the dimensions of texts and images are important factors for multi-modal data retrieval, we set two kinds of different dimensions for comparisons: one is 500-dimension BoVW and 1000-dimension BoW, the other is 1000-dimension BoVW and 5000-dimension BoW.

The other data set we used is the NUS-WIDE data set. Each image with its annotated tags in NUS-WIDE can be taken as a pair of image-text data. We only select those pairs that belong to one of the 10 largest classes with each pair exclusively belonging to one of the 10 classes. We use the 500-dimension BoVW based on SIFT features for the representation of each image and 1000-dimension tags for the representation of each text as the authors supplied.

## Evaluation Methods

There are many evaluation criteria for cross-modal retrieval algorithms such as mean average precision (MAP), area under curve (AUC) and precision recall curves. Most of them are based on the retrieved ranking list of queries. Ideally, given labeled pairs of image-text, an appropriately correct retrieved result can be one that belongs to the same category as the query data (Sharma et al. 2012) or the corresponding unique one paired with the query (Jia, Salzmann, and Darrell 2011). The first one represents the ability of learning discriminative cross-modal mapping functions while the later one reveals the ability of learning corresponding latent concepts. In this paper, we use both of them as follows:

**MAP** : MAP is defined here to measure whether the retrieved data belong to the same class as the query (*relevant*) or does not belong to the same class (*irrelevant*). Given a query (one image or one text) and a set of its corresponding $R$ retrieved data, the Average Precision is defined as

$$AP = \frac{1}{L}\sum_{r=1}^{R} prec(r)\delta(r), \tag{13}$$

where L is the number of relevant data in the retrieved set, $prec(r)$ represents the precision of the $r$ retrieved data. $\delta(r) = 1$ if the $rth$ retrieved datum is relevant to the query and $\delta(r) = 0$ otherwise. MAP is defined as the average AP of all the queries. Same as (Zhen and Yeung 2012), we set R = 50 in the experiments.

**Percentage**: Since there is only one ground-truth match for each image/text, to evaluate the multi-modal performance we can resort to the position of the ground-truth textt/image in the ranked list obtained. In general, one image (or text) is considered correctly retrieved if it appears in the first $t$ percent of the ranked list of its corresponding

| NUS-WIDE | Image Query Text | Text Query Image |
|---|---|---|
| CCA | 0.2175 | 0.2400 |
| GMA | 0.2634 | **0.3051** |
| SCDL | 0.3073 | 0.2602 |
| SliM$^2$ | **0.3154** | 0.2924 |

Table 3: The performance comparison in terms of MAP scores on NUS-WIDE data set. The results shown in boldface are best results.

| NUS-WIDE | Image Query Text | Text Query Image |
|---|---|---|
| CCA | 0.3901 | **0.4016** |
| GMA | 0.4242 | 0.2913 |
| SCDL | 0.4421 | 0.3239 |
| SliM$^2$ | **0.4639** | 0.3877 |

Table 4: The performance comparison in terms of Percentage scores on NUS-WIDE data set. The results shown in boldface are best results.

retrieved texts (or images) according to (Jia, Salzmann, and Darrell 2011). $t$ is set to equal to 0.2 in our experiments.

### Compared Methods

We devise our compared algorithms as follows : compare with one of the popular traditional methods only utilizing the pair-wise information, one of our counterparts and the unsupervised dictionary learning method with a mapping function cross reconstruction coefficients. The compared algorithms with our proposed SliM$^2$ are listed as follows:

- **Canonical Correlational Analysis (CCA)**: CCA is the classical method in cross modal retrieval which learns a common space across multi-modal data.
- **Generalized Multiview Analysis (GMA)**: GMA is a supervised method in cross-modal retrieval which utilizes both pair-wised and label information of multi-modal data. As stated by authors (Sharma et al. 2012), GMA is a supervised kernelizable extension of CCA and maps data in different modality spaces to a single (non) linear subspace.
- **Semi-coupled Dictionary Learning (SCDL)**: SCDL (Wang et al. 2012) is an unsupervised dictionary learning approach to learn a pair of dictionaries and a mapping function across two-views in image domains, here we conduct SCDL to multi-modal data.

### Parameter Tuning

For parameter tuning, we split the training data sets into 5 folds and test on each fold with the remaining 4 as training data to do cross validation. $\beta, \gamma, \lambda_m (m \in \{1, 2\})$ and $K$ are tuning parameters in our experiments. We perform grid search strategy on the first 4 folds to set $\lambda_m (m \in \{1, 2\})$ and line-search method for the other parameters. The setting of $\beta, \gamma, \lambda_1, \lambda_2$ and $K$ on Wiki data set is $1, 0.1, 0.1, 0.01$ and $200$, respectively while $0.01, 1, 0.01, 0.01$ and $128$ on NUS-WIDE data set. Here, $\lambda_1$ is the regularization parameter corresponding to image modality while $\lambda_2$ corresponds to text modality.

### Performance Comparisons

For the Wiki Text-Image data set, the performance by each algorithm is given in table 1 and table 2 in terms of MAP and Percentage respectively. For NUS-WIDE data, the performance by each algorithm is given in table 3 and table 4 in terms of MAP and Percentage respectively.

In our experiments, we can submit one image to retrieve texts (Image query Text), or submit one text to retrieve images (Text query Image). From the experiments, we can make the following observations:

- *For Image query Text*, in general, dictionary learning based methods (SCDL and SliM$^2$) are better than direct mapping-based methods (CCA and GMA) on image query text case in all of metrics for the two data sets, and moreover SliM$^2$ achieves the best performances. This is due to that SCDL and SliM$^2$ learn the multi-modal mapping functions from sparse codes instead of BoW/BoVW with sparse codes obtaining through the minimization of reconstruction errors. The introduction of class label further boosts the multi-modal retrieval.
- *For Text query Image*, the proposed SliM$^2$ achieves best performances in term of Percentage metric over Wiki data set. Since images and texts are paired in our experiments, Percentage is more accurate for true performance. CCA shows a good performance over NUS-WIDE data set for percentage because the annotated tags in NUS-WIDE are manually selected and there is highly-qualified correlation between images and tags.
- *For different algorithms*, the algorithms utilize pair-wise information perform better on Percentage with algorithms utilized label information better on MAP.

Figure 1 illustrates one example of image query text and one example of text query image over Wiki image-text data set. The retrieved results by SliM$^2$ (top row) and GMA (bottom row) are compared.

For the example of image query text, we use the corresponding images of retrieved texts to demonstrate the results. Though all of retrieved texts come from the ”sports” category same as the query image, and strongly correspond to the query image, the result by SliM$^2$ is more visually consistent with the query image.

For the example of text query image, the query text is about parks from ”geography” category. The retrieved images by SliM$^2$ all come from ”geography” category, while the first retrieved image and the last one by GMA come from ”history” category. From the underlined words in the query text describing the semantics of this query text, we can observe that the retrieved images by SliM$^2$ are more semantically correlated with the query text than that of GMA.

## Conclusion

SliM$^2$ is proposed in this paper for multi-modal retrieval. SliM$^2$ can utilize the class information to jointly learn discriminative multi-modal dictionaries as well as mapping functions between different modalities. We have demonstrated the superior performance of SliM$^2$ in terms of MAP and Percentage for two data sets.

| Wiki | BoVW(500$D$),BoW(1000$D$) | | BoVW(1000$D$),BoW(5000$D$) | |
|---|---|---|---|---|
| | Image Query Text | Text query Image | Image Query Text | Text query Image |
| CCA | 0.1767 | 0.1809 | 0.1994 | 0.1859 |
| GMA | 0.2245 | **0.2148** | 0.2093 | **0.2267** |
| SCDL | 0.2341 | 0.1988 | 0.2527 | 0.1981 |
| SliM$^2$ | **0.2399** | 0.2025 | **0.2548** | 0.2021 |

Table 1: The performance comparison in terms of MAP scores on Wiki data set. 500-dimensional bag of visual words (BoVW) and 1000-dimensional bag of textual words (BoW), as well as 1000-dimensional bag of visual words (BoVW) and 5000-dimensional bag of textual words (BoW), are used to represent each image and text respectively. The results shown in boldface are best results.

| Wiki | BoVW(500$D$),BoW(1000$D$) | | BoVW(1000$D$),BoW(5000$D$) | |
|---|---|---|---|---|
| | Image Query Text | Text Query Image | Image Query Text | Text Query Image |
| CCA | 0.2236 | 0.2340 | 0.3054 | 0.2845 |
| GMA | 0.2877 | 0.2548 | 0.3002 | 0.2496 |
| SCDL | 0.3709 | 0.2790 | 0.3857 | 0.3037 |
| SliM$^2$ | **0.3899** | **0.2842** | **0.4084** | **0.3106** |

Table 2: The performance comparison in terms of Percentage scores on Wiki data set. 500-dimensional bag of visual words (BoVW) and 1000-dimensional bag of textual words (BoW), as well as 1000-dimensional bag of visual words (BoVW) and 5000-dimensional bag of textual words (BoW), are used to represent each image and text respectively. The results shown in boldface are best results.

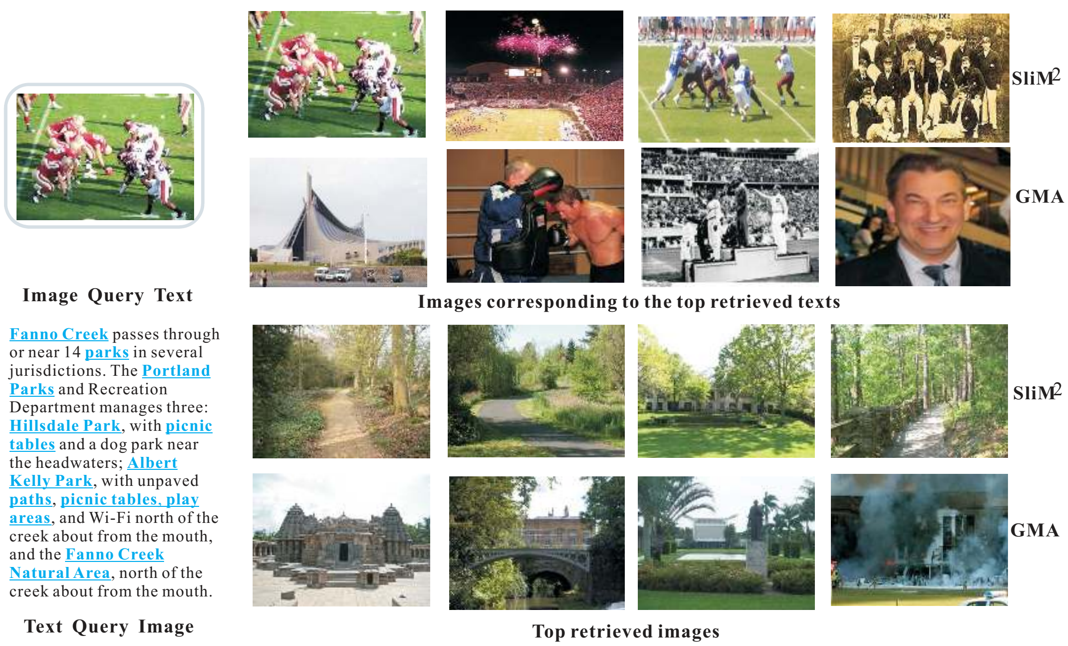


Figure 1: Two examples of image query text and text query image over Wiki data set by SliM$^2$ (top row) and GMA (bottom row). For the example of image query text, we use the corresponding images of retrieved texts to demonstrate the results. The query image comes from the ”sports” category and all of retrieved texts (and their corresponding images) also come from ”sports” category. For the example of image query text, the query text is about parks from ”geography” category. The underlined words in the query text describe the semantics of the query text. All of retrieved images by SliM$^2$ come from ”geography” category, and the second and the third retrieved images by GMA come from ”geography” category while the other two come from ”history” category.

## Acknowledgements

This work is supported by 973 Program (No. 2012CB316400), NSFC (61070068, 90920303) and 863 program (2012AA012505).

## References


Bengio, S.; Pereira, F.; Singer, Y.; and Strelow, D. 2009. Group sparse coding. *Advances in Neural Information Processing Systems* 22:82–89.

Blei, D., and Jordan, M. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 127–134. ACM.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, 178–178. IEEE.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika* 28(3/4):321–377.

Jia, Y.; Salzmann, M.; and Darrell, T. 2010. Factorized latent spaces with structured sparsity. *Advances in Neural Information Processing Systems* 23:982–990.

Jia, Y.; Salzmann, M.; and Darrell, T. 2011. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2407–2414. IEEE.

Jia, K.; Tang, X.; and Wang, X. 2012. Image transformation based on learning dictionaries across image spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kang, Z.; Grauman, K.; and Sha, F. 2011. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning*, 521–528.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, 1150–1157. Ieee.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Supervised dictionary learning. *arXiv preprint arXiv:0809.3083*.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research* 11:19–60.

Monaci, G.; Jost, P.; Vandergheynst, P.; Mailhe, B.; Lesage, S.; and Gribonval, R. 2007. Learning multimodal dictionaries. *Image Processing, IEEE Transactions on* 16(9):2272–2283.

Olshausen, B.; Field, D.; et al. 1997. Sparse coding with an overcomplete basis set: A strategy employed by vi? *Vision research* 37(23):3311–3326.

Putthividhy, D.; Attias, H.; and Nagarajan, S. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 3408–3415. IEEE.

Qin, Z.; Scheinberg, K.; and Goldfarb, D. 2010. Efficient block-coordinate descent algorithms for the group lasso. *Preprint*.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, 251–260. ACM.

Salomatin, K.; Yang, Y.; and Lad, A. 2009. Multi-field correlated topic modeling. *SDM09* 628–637.

Sharma, A.; Kumar, A.; Daume, H.; and Jacobs, D. 2012. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2160–2167. IEEE.

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

Todorovic, S., and Nechyba, M. 2004. Detection of artificial structures in natural-scene images using dynamic trees. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, 35–39. IEEE.

Virtanen, S.; Jia, Y.; Klami, A.; and Darrell, T. 2012. Factorized multi-modal topic model. *arXiv preprint arXiv:1210.4920*.

Wang, S.; Zhang, L.; Liang, Y.; and Pan, Q. 2012. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2216–2223. IEEE.

Wright, J.; Yang, A.; Ganesh, A.; Sastry, S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(2):210–227.

Wu, F.; Zhang, H.; and Zhuang, Y. 2006. Learning semantic correlations for cross-media retrieval. In *Image Processing, 2006 IEEE International Conference on*, 1465–1468. IEEE.

Yang, M.; Zhang, L.; Yang, J.; and Zhang, D. 2010. Metaface learning for sparse representation based face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 1601–1604. IEEE.

Zhen, Y., and Yeung, D.-Y. 2012. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 940–948. ACM.

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