Self-Paced Cross-Modal Subspace Matching

Jian Liang[1], Zhihang Li[1], Dong Cao[1], Ran He[1,2]*, Jingdong Wang[3]
[1] Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences (CAS), Beijing 100190, China
[2] Center for Excellence in Brain Science and Intelligence Technology, CAS, Shanghai 200031, China
[3] Microsoft Research, Beijing 100190, China
{jian.liang, zhihang.li, dong.cao, rhe}@nlpr.ia.ac.cn, jingdw@microsoft.com

ABSTRACT

Cross-modal matching methods match data from different modalities according to their similarities. Most existing methods utilize label information to reduce the semantic gap between different modalities. However, it is usually time-consuming to manually label large-scale data. This paper proposes a Self-Paced Cross-Modal Subspace Matching (SCSM) method for unsupervised multimodal data. We assume that multimodal data are pair-wised and from several semantic groups, which form hard pair-wised constraints and soft semantic group constraints respectively. Then, we formulate the unsupervised cross-modal matching problem as a non-convex joint feature learning and data grouping problem. Self-paced learning, which learns samples from 'easy' to 'complex', is further introduced to refine the grouping result. Moreover, a multimodal graph is constructed to preserve the relationship of both inter- and intra-modality similarity. An alternating minimization method is employed to minimize the non-convex optimization problem, followed by the discussion on its convergence analysis and computational complexity. Experimental results on four multimodal databases show that SCSM outperforms state-of-the-art cross-modal subspace learning methods.

Keywords

Cross-Modal Matching; Heterogeneous Data; Unsupervised Subspace Learning; Self-Paced Learning

1. INTRODUCTION

Nowadays, multimodal data spring up as the popularity of social network on the Internet. People would like to upload personalized content through various forms such as text, image, audio and video simultaneously, this phenomenon urgently pushes the need for heterogeneous content retrieval. Compared with unimodal retrieval, cross-modal retrieval exploiting heterogeneous feature representations to discover optimal heterogeneous content for a given user query, is more and more frequently utilized. Although search engines also provide such services, e.g., text-image search, they mainly rely on homogeneous contents associated with heterogeneous contents within the same web pages or tweets. The main challenge in cross-modal matching is how to bridge the semantic gap between heterogeneous modalities (e.g., different distributions, different dimensionalities). In this literature, some promising approaches [31, 37, 44] have been developed to alleviate this gap. Probability based learning algorithms [28, 16] along with subspace based learning algorithms [12, 29, 37, 18, 22] are even more favored and widely applied for cross-modal learning applications, e.g., image annotation and cross-language information retrieval.

Topic models, such as Latent Dirichlet Allocation (LDA) [6], have proved effective at describing the underlying topics in one single modality. Following researches [5, 28] focused on correlating topics among different modalities, and extended it to the multi-modal case. While [5] assumed a one-to-one correspondence between the topics of each modality, and [28] assumed a regression module between two sets of topics instead of one-to-one correspondence, and achieved better performance with more freedom allowed. [16] further defined a Markov random field to capture the relationships between different topic sets in every modality. Besides traditional topic models, [35] adopted deep Boltzmann machines to learn a generative model, and [27] proposed an effective nonparametric Bayesian framework based on the Indian Buffet Process (IBP) for integrating multimodal data in a latent space. Besides, some recent work [1, 19] exploited deep models for cross-modal matching. And [38] further investigated the impact of deep features on cross-modal retrieval.

In contrast, subspace based cross-modal methods seem attractive due to its efficiency. They aimed to learn a latent common subspace to make all modalities of data to be close to each other. In this literature, Canonical Correlation Analysis (CCA) [12] is the most popular one to obtain such a common space. CCA tries to find two linear projections to maximally preserve the mutual correlations among multimodal data. Owing to its simplicity and effectiveness, CCA has been widely applied to the cross-modal retrieval [7], face recognition [21] and word embedding [8, 9]. Another popular method is the Partial Least Squares (PLS) [31] that learns orthogonal score vectors by maximizing the covariance between different multimodal data. Although both CCA and PLS are able to solve cross-modal subspace matching, the valuable label information

[*] Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2911451.2911527

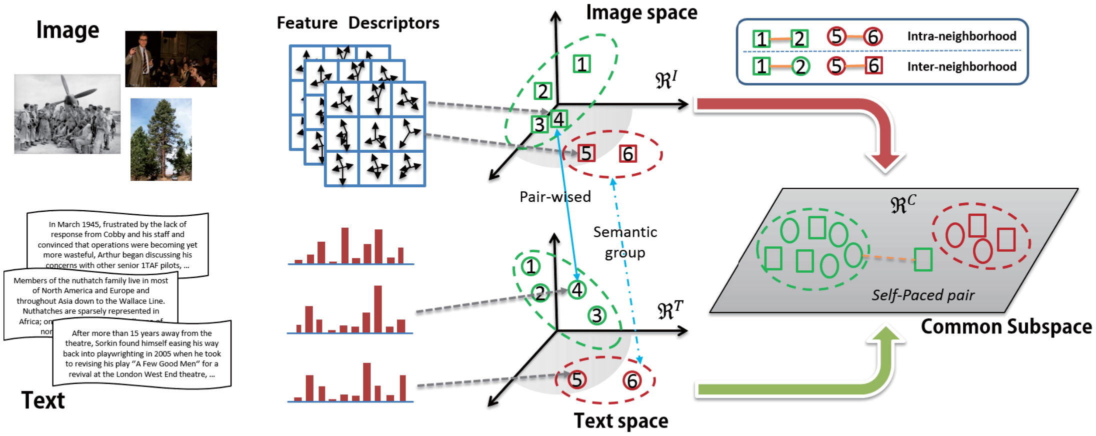


Image
Feature Descriptors
Image space
$\Re^I$
1
2
3
4
5
6
Intra-neighborhood
Inter-neighborhood
Pair-wised
Semantic group
$\Re^C$
Self-Paced pair
Common Subspace
$\Re^T$
In March 1945, frustrated by the lack of response from Cobby and his staff and convinced that operations were becoming yet more wasteful, Arthur began discussing his concerns with other senior 1TAF pilots, ...
Members of the nuthatch family live in most of North America and Europe and throughout Asia down to the Wallace Line. Nuthatches are sparsely represented in Africa; on
After more than 15 years away from the theatre, Sorkin found himself easing his way back into playwrighting in 2005 when he took to revising his play "A Few Good Men" for a revival at the London West End theatre, ...
Text
Text space

Figure 1: Overview of the proposed SCSM for unsupervised cross-modal matching. Handling hard pair-wised constraints and soft semantic group constraints (labels are unknown) simultaneously, SCSM is formulated as a non-convex joint feature learning and data grouping problem. Self-paced learning is further introduced to refine the grouping result so that we can train a model from 'easy' pairs to 'complex' pairs. Moreover, a novel multimodal similarity graph is leveraged to preserve the local intra- and inter- neighborhoods in the common subspace $\mathcal{R}^C$. (Best viewed in colors).

has not been fully utilized to reduce the semantic gap between text and image modalities.

Since label information potentially reduces the semantic gap between the low-level image features and high-level document descriptions, supervised methods have drawn considerable attentions in cross-modal subspace learning. A generalized multi-view framework [33] was presented to learn a discriminative common space. Built on CCA, [11] proposed to incorporate the third view (label information) to capture the high-level semantics. To select relevant and discriminative features simultaneously, [37] developed a coupled linear regression framework. Moreover, the authors of [13] seek the common structure hidden in heterogeneous modalities via the pairwise constraint. Hashing methods have been also used successfully in multimodal problem [24, 39] and cross-modal problem [40, 43, 41]. While [40] utilized discriminative coupled dictionaries to learn better hash functions, [43] exploited matrix factorization to learn the hash functions. Despite the improvement brought by these supervised cross-modal methods, they will suffer from unlabeled data exceedingly.

To address these above problems, we propose a self-paced joint learning framework (as shown in Figure 1) for the unsupervised cross-modal matching problem by considering feature learning and data grouping simultaneously, which result in a non-convex objective. First, we assume that multimodal data are pair-wised and from several semantic groups, which form hard pair-wised constraints and soft semantic group constraints respectively. For these hard pair-wised constraints, two linear projections are learned to map different modalities into a common space respectively. For the soft semantic group constraints, self-paced learning strategy [20, 42] is applied to refine the grouping results so that we can train a model from 'easy' pairs to 'complex' pairs. Second, the projection regularization is employed for feature learning. A graph-based regularization term preserves the relationship of both inter- and intra- modality similarities. Finally, we present an alternating algorithm to cope with this optimization problem. Extensive experimental results on four benchmark databases demonstrate that our approach is highly effective.

The contributions of this paper are summarized as follows:

1) We propose a general framework for unsupervised cross-modal subspace learning that can handle hard pair-wised constraints and soft semantic group constraints simultaneously. A joint feature learning and data grouping formulation is accordingly developed and results in a non-convex problem.

2) Self-paced learning is used to alleviate the inaccurate estimation of soft semantic group so that we can gradually include sample pair sequences from 'easy' to 'complex'. Multimodal graph, preserving the inter-modality and intra-modality similarity, is then utilized to better explore unsupervised multimodal data.

3) An alternating minimization method is put forward to efficiently minimize the proposed non-convex optimization problem. Experimental results validate that our method outperforms state-of-the-art unsupervised subspace learning methods or even better than some supervised ones.

The remainder of the paper is organized as follows. In Section 2, we will illustrate the details of our SCSM method. Experimental results are presented in Section 3. Finally, the conclusion is summarized in Section 4.

2. SELF-PACED CROSS-MODAL SUBSPACE MATCHING

In this section, we will present our SCSM in details and use two modalities for example in this paper. Note that SCSM can be easily extended to cover the case for more than two modalities.

2.1 Notation and Problem Definition

We start with a brief introduction to some notations. For a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, its i-th row, j-th column are denoted by $\mathbf{m}_i$, $\mathbf{m}^j$ respectively, and $M_{i,j}$ lies in the i-th row and j-th column. The Frobenius norm of the matrix $\mathbf{M}$ is defined as $||\mathbf{M}||_F = \sqrt{\sum_{i=1}^{n} ||\mathbf{m}_i||_2^2}$, and the trace of the square matrix M is defined as $Tr(M) = \sum_i M_{i,i}$.

Assume that we have two sets of features from different modalities (e.g., image and text), $\mathbf{X}_a = [\mathbf{x}_1^a, \mathbf{x}_2^a, \dots, \mathbf{x}_n^a] \in \mathbb{R}^{d_1 \times n}$, $\mathbf{X}_b = [\mathbf{x}_1^b, \mathbf{x}_2^b, \dots, \mathbf{x}_n^b] \in \mathbb{R}^{d_2 \times n}$, where d_i is the dimensionality of the i-th modality, n is the amount of training image-text pairs. And each pair $\{\mathbf{x}_i^a, \mathbf{x}_i^b\}$ has the same underlying content and belongs to the same class, i.e., the hard pair-wised constraint. However, the concrete label of each pair is unknown here.

The goal of cross-modal matching is to obtain two subspaces $\mathbf{S}_p = \mathbf{U}_p^T \mathbf{X}_p \in \mathbb{R}^{c \times n}, p \in \{a, b\}$ with the same dimensionality, where $\mathbf{U}_p \in \mathbb{R}^{d_p \times c}, p \in \{a, b\}$ denote two projection matrices for these two modalities $\mathbf{X}_p, p \in \{a, b\}$, respectively. Cross-modal matching tasks usually include: 1) using texts to match the related images, and 2) using images to match the related texts.

Here, we mainly focus on the case that one retrieved result is a good matching only when it shares the same semantic label with the given user query.

2.2 Self-Paced Learning Revisit

Curriculum learning [4] and self-paced learning [20] have been attracting increasing attention in machine learning, computer vision and multimedia analysis fields. The philosophy under these concepts is to simulate the learning process of humans/animals, i.e., humans/animals generally start with learning easier aspects of a learning task, and then gradually take more complex examples into consideration [25]. Instead of using the aforementioned heuristic strategies in curriculum learning, self-paced learning automatically includes training samples from 'easy' to 'complex' in a purely self-paced way.

Given a training dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^n$, in which $\{x_i, y_i\}$ denotes the i-th observed sample and its label, then let $L(y_i, g(x_i, w))$ be the loss function, w is the model parameter inside the decision function g. Generally, the objective of self-paced learning consists of two parts, i.e., a weighted loss term on all samples and a general self-paced regularizer imposed on sample weights, is expressed as:

$$\min_{w, v \in [0,1]^n} E(w, v; \lambda) = \sum_{i=1}^n (v_i L(y_i, g(x_i, w)) + f(v_i, \lambda)), \quad (1)$$

where λ is the age parameter for controlling the learning rate, and $f(v, \lambda)$ is the self-paced regularizer. A formal definition is given in [17, 42]. This strategy, as supported by empirical evaluation, has proved helpful in alleviating the local optimum problem in non-convex optimization [3].

2.3 Model Formulation

In Figure 1, multimodal data are firstly pair-wised, further supposed to come from several latent semantic groups, which form hard pair-wised constraints and soft semantic group constraints respectively.

Latent Discriminative Subspace Learning For the cross-modal problem, a CCA-like objective function expressed as

$$\min_{\mathbf{U}_a, \mathbf{U}_b} \|\mathbf{U}_a^T \mathbf{X}_a - \mathbf{U}_b^T \mathbf{X}_b\|_F^2 + \Phi(\mathbf{U}_a, \mathbf{U}_b), \quad (2)$$

where $\Phi(\cdot)$ is a regularizer imposed on the projection matrices, is exploited frequently to discover the optimal common subspace. Then we consider the following objective,

$$\min_{\mathbf{U}_a, \mathbf{U}_b, \mathbf{Y}} \sum_{p \in \{a,b\}} \|\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y}\|_F^2 + \Phi(\mathbf{U}_a, \mathbf{U}_b), \quad (3)$$

as a relaxed candidate for that in Eq.(2), owing to the basic inequality $\|A - C\|_F^2 + \|B - C\|_F^2 \geq \frac{\|A-B\|_F^2}{2}$.

Since multimodal data are separated into several groups in the latent subspace, we further strict the latent variable $\mathbf{Y}$ lying in the discrete space $\{0, 1\}^{c \times n}$, revealing the group/cluster membership, where c is the amount of latent groups. Obviously, $\mathbf{1}_c \mathbf{Y} = \mathbf{1}_n$, where $\mathbf{1}_c$ and $\mathbf{1}_n$ denote the constant vectors of all one, of dimensions c and n respectively. Let $\Phi(\mathbf{U}_a, \mathbf{U}_b) = \beta \left(\|\mathbf{U}_a\|_F^2 + \|\mathbf{U}_b\|_F^2\right)$ as [2], we can write the multimodal feature learning term as follows:

$$\begin{aligned} &\min_{\mathbf{U}_a, \mathbf{U}_b, \mathbf{Y}} \sum_{p \in \{a,b\}} \|\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y}\|_F^2 + \beta \|\mathbf{U}_p\|_F^2, \\ &s.t. \quad \mathbf{Y} \in \{0, 1\}^{c \times n}, \sum_i^c Y_{i,j} = 1, \forall j \in [1, n], \end{aligned} \quad (4)$$

where each $\mathbf{X}_p$ is the feature representation matrix, $\mathbf{U}_p$ is the corresponding projection matrix, and $\|\mathbf{U}_p\|_F^2$ is a regularizer imposed on project matrices to avoid trivial solutions.

Cluster indicator $\mathbf{Y}$ is a discrete variable, which easily involves the optimization method in a local optimum. Fortunately, the learning process from 'easy' to 'complex' proposed in self-paced learning can effectively avoid the local optimum. Thus we incorporate self-paced learning model in Eq.(1) into Eq.(4), resulting in the following form:

$$\begin{aligned} &\min_{\mathbf{U}_a, \mathbf{U}_b, \mathbf{v}, \mathbf{Y}} \sum_{p \in \{a,b\}} \sum_{i=1}^n v_i \ell_i + \beta \sum_{p \in \{a,b\}} ||\mathbf{U}_p||_F^2 + f(\mathbf{v}; k) \\ &s.t. \quad \mathbf{Y} \in \{0, 1\}^{c \times n}, \sum_i^c Y_{i,j} = 1, \forall j \in [1, n], \end{aligned} \quad (5)$$

where the i-th loss of sample is defined as $\ell_i = ||(\mathbf{U}_p^T \mathbf{x}_p^i - \mathbf{y}^i)||_F^2$ above.

Multimodal Locality Preserving The term in Eq.(4) mainly focuses on the similarities among pair-wised heterogeneous data. Local similarity in each modality may vanish in the process of feature learning. We then take locality preserving into consideration, which has also proven effective in [14], the projected features in the common subspace $\mathcal{R}^C$ should inherit the similar local structure to it in these two original feature spaces. Then for each projected modality, we try to minimize the following objective function:

$$\sum_{i,j} (z_p^i - z_p^j)^2 \mathbf{W}_{ij}, \quad (6)$$

where $Z_p = \mathbf{U}_p^T \mathbf{X}_p$ are the projected feature matrices, and W_{ij} is the intra-similarity between the i-th and the j-th sample. We adopt the Gaussian kernel function $d(\mathbf{x}_p^i, \mathbf{x}_p^j) = e^{\frac{-||\mathbf{x}_p^i - \mathbf{x}_p^j||^2}{2\sigma^2}}$ to measure the local similarity:

$$W_{ij}^p = \begin{cases} d(\mathbf{x}_p^i, \mathbf{x}_p^j) & \text{if } \mathbf{x}_p^i \in N_r(\mathbf{x}_p^j) \text{ or } \mathbf{x}_p^j \in N_r(\mathbf{x}_p^i), \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where $N_r(\cdot)$ is the set of r-nearest samples. The total intra-similarity preserving term in Eq.(6) for each modality can be formulated as:

$$\mathcal{L}_{intra} = \sum_{p \in \{a,b\}} Tr(\mathbf{Z}_p \mathbf{L}_p \mathbf{Z}_p^T), \quad (8)$$

where $\mathbf{L}_p = \mathbf{D}^p - \mathbf{W}^p$ is the Laplacian matrix and $\mathbf{D}^p$ is a diagonal matrix, where $D_{i,i}^p = \sum_j W_{i,j}^p$.

Moreover, a novel similarity preserving term is introduced to maximize the similarity between data from one identical latent cluster. Given the cluster indicator $\mathbf{Y}$ instead of the missing semantic labels in [36], $W^{ab} = W^{ba} = \mathbf{Y}^T\mathbf{Y}$ can be seen as the inter-similarity matrix. Apparently, only paired data from the same cluster are treated as similar pair, the similarity of coupled data from different clusters equals to 0, otherwise. We can define a inter-similarity preserving term akin to the intra-similarity term above, expressed as:

$$\mathcal{L}_{inter} = Tr(\mathbf{Z}_a \mathbf{L}_{ab} \mathbf{Z}_b^T). \tag{9}$$

Then we introduce a new joint similarity matrix as below:

$$\mathbf{W} = \begin{bmatrix} \gamma W^a & W^{ab} \\ W^{ba} & \gamma W^b \end{bmatrix}. \tag{10}$$

Combining these two locality-preserving terms $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ together, we further obtain the following term:

$$\begin{aligned} \mathcal{L}(\mathbf{U}_a, \mathbf{U}_b) &= \mathcal{L}_{inter} + \gamma \mathcal{L}_{intra} \\ &= \sum_{p,q \in \{a,b\}} Tr(\mathbf{U}_p^T \mathbf{X}_p \mathbf{L}_{pq} \mathbf{X}_q^T \mathbf{U}_q), \end{aligned} \tag{11}$$

where $\mathbf{L}_{pq}$ is the corresponding block of the Laplacian matrix $\mathbf{L}$, and γ is a trade-off parameter to balance the intra- and inter- similarity preserving terms.

Associating discriminative subspace learning with local similarity preserving together, we obtain the overall objective function for SCSM:

$$\begin{aligned} \min_{\mathbf{U}\{a,b\}, \mathbf{v}, \mathbf{Y}} & \sum_{p \in \{a,b\}} \sum_{i=1}^{n} v_i \ell_i + \alpha \mathcal{L}(\mathbf{U}_a, \mathbf{U}_b) \\ & + \beta \sum_{p \in \{a,b\}} ||\mathbf{U}_p||_F^2 + f(\mathbf{v}; k) \\ s.t. \quad & \mathbf{Y} \in \{0,1\}^{c \times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n]. \end{aligned} \tag{12}$$

2.4 Optimization Algorithm

In this subsection, a fast iterative optimization algorithm is developed to find the optimal solution of Eq.(12). Regarding the self-paced regualrizer $f(\cdot)$, we adopt the same strategy utilized in [20]: $f(\mathbf{v}; k) = -\frac{1}{k}\sum_i v_i$, where $\mathbf{v} \in \mathbb{R}^n$ denotes the weights imposed on the loss term, which is a binary vector. When age parameter k is given in each iteration, the current indicator variable v_i can be defined for each sample x_i as:

$$v_i = \begin{cases} 1 & \text{if } \ell_i \leq \frac{1}{k}, \\ 0 & \text{if } \ell_i > \frac{1}{k}. \end{cases} \tag{13}$$

Besides, we rewrite Eq.(12) into the following form:

$$\begin{aligned} \min_{\mathbf{U}\{a,b\}, \mathbf{v}, \mathbf{Y}} & \sum_{p \in \{a,b\}} ||(\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y}) diag(\mathbf{v})||_F^2 \\ & + \alpha \sum_{p \in \{a,b\}} Tr(\mathbf{U}_p^T \mathbf{X}_p \mathbf{L}_{pq} \mathbf{X}_q^T \mathbf{U}_q) \\ & + \beta \sum_{p \in \{a,b\}} ||\mathbf{U}_p||_F^2 - \frac{1}{k} \sum_i v_i \\ s.t. \quad & \mathbf{Y} \in \{0,1\}^{c \times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n]. \end{aligned} \tag{14}$$

Since the matrix $\mathbf{Y} \in \{0,1\}^{c \times n}$ is limited to discrete values, i.e., a non-convex set, Eq.(14) is a non-convex problem. We employ an alternating minimization algorithm for Eq.(14), the involved parameters include $\mathbf{U}_a, \mathbf{U}_b, \mathbf{v}$ and $\mathbf{Y}$.

1) Solve $\mathbf{U}_a, \mathbf{U}_b$ *when* $\mathbf{Y}, \mathbf{v}$ *are fixed.* Optimizing the objective function(14) is equal to minimizing the following problem:

$$\begin{aligned} \min_{\mathbf{U}_p} & ||(\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y}) \mathbf{V}||_F^2 \\ & + \alpha \sum_p \sum_q Tr(\mathbf{U}_p^T \mathbf{X}_p \mathbf{L}_{pq} \mathbf{X}_q^T \mathbf{U}_q) + \beta ||\mathbf{U}_p||_F^2, \end{aligned} \tag{15}$$

where $\mathbf{V}$ is a diagonal matrix whose entries correspond to the elements in v. Differentiating the objective function in Eq.(15) with respect to $\mathbf{U}_p$ and setting it to zero, we have the following equation:

$$\begin{aligned} & (\mathbf{X}_p \mathbf{V} \mathbf{V}^T \mathbf{X}_p^T + \alpha \mathbf{X}_p \mathbf{L}_{pp} \mathbf{X}_p^T + \beta \mathbf{I}) \mathbf{U}_p \\ & \quad = \mathbf{X}_p \mathbf{V} \mathbf{V}^T \mathbf{Y}^T - \alpha \mathbf{X}_p \mathbf{L}_{pq} \mathbf{X}_q^T \mathbf{U}_q \end{aligned} \tag{16}$$

Then, the optimal solution of Eq.(16) can be computed via solving the above linear system problem.

2) Solve $\mathbf{Y}$ *when* $\mathbf{U}_a, \mathbf{U}_b, \mathbf{v}$ *are fixed.* Optimizing the objective function in Eq.(14) is equal to optimizing the following problem:

$$\begin{aligned} \min_{\mathbf{Y}} & \sum_{p \in \{a,b\}} ||(\mathbf{U}_p^T \mathbf{X}_p - \mathbf{Y}) \mathbf{V}||_F^2 + 2\alpha Tr(\mathbf{U}_a^T \mathbf{X}_a \mathbf{Y}^T \mathbf{Y} \mathbf{X}_b^T \mathbf{U}_b) \\ s.t. \quad & \mathbf{Y} \in \{0,1\}^{c \times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n]. \end{aligned} \tag{17}$$

It is challenging to directly minimize the above objective function w.r.t $\mathbf{Y}$ due to the discrete constraints, resulting in a NP hard problem. Inspired by [34], we can optimize $\mathbf{Y}$ column by column, i.e., optimize one column of $\mathbf{Y}$ with all the other columns fixed. So we can iteratively learn one column at one time. Then Eq.(17) is equivalent to:

$$\begin{aligned} \min_{\mathbf{Y}} \quad & Tr(\mathbf{V}^T \mathbf{Y}^T \mathbf{Y} \mathbf{V}) + \alpha Tr(\mathbf{E} \mathbf{Y}^T \mathbf{Y} \mathbf{F}^T) \\ & - Tr(\mathbf{G} \mathbf{Y}^T) - Tr(\mathbf{H} \mathbf{Y}^T) \\ & s.t. \quad \mathbf{Y} \in \{0,1\}^{c \times n}, \sum_{i}^{c} Y_{i,j} = 1, \forall j \in [1, n], \end{aligned} \tag{18}$$

where $\mathbf{E} = \mathbf{U}_a^T \mathbf{X}_a$, $\mathbf{F} = \mathbf{U}_b^T \mathbf{X}_b$, $\mathbf{G} = \mathbf{U}_a^T \mathbf{X}_a \mathbf{V} \mathbf{V}^T$, $\mathbf{H} = \mathbf{U}_b^T \mathbf{X}_b \mathbf{V} \mathbf{V}^T$.

The objective in Eq.(18) can be further attributed to minimizing two following general cases, i.e., $Tr(\mathbf{A}\mathbf{Y}^T\mathbf{Y}\mathbf{B}^T)$ and $Tr(\mathbf{C}\mathbf{Y}^T)$. Let $\mathbf{y}$ be the i-th column of $\mathbf{Y}$, $i = 1, \cdots, n$, and $\widetilde{\mathbf{Y}}$ the matrix $\mathbf{Y}$ of excluding $\mathbf{y}$ where $\mathbf{y}$ is one of all n samples. Similarly, denote by $\mathbf{a}$ the i-th column of $\mathbf{A}$, $\widetilde{\mathbf{A}}$ the matrix of $\mathbf{A}$ excluding $\mathbf{a}$, $\mathbf{b}$ the i-th column of $\mathbf{B}$, and $\widetilde{\mathbf{B}}$ the matrix of $\mathbf{B}$ excluding $\mathbf{b}$ and $\mathbf{c}$ the i-th column of $\mathbf{C}$ and $\widetilde{\mathbf{C}}$ the matrix of $\mathbf{C}$ excluding $\mathbf{c}$. Then we have the following form:

$$\begin{aligned} Tr(\mathbf{A}\mathbf{Y}^T\mathbf{Y}\mathbf{B}^T) &= Tr((\mathbf{a}\mathbf{y}^T + \widetilde{\mathbf{A}}\widetilde{\mathbf{Y}}^T)(\mathbf{y}\mathbf{b}^T + \widetilde{\mathbf{Y}}\widetilde{\mathbf{B}}^T)) \\ &= const + Tr(\mathbf{a}\mathbf{y}^T\mathbf{y}\mathbf{b}^T) + \mathbf{y}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{B}}^T\mathbf{a} + \mathbf{y}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{A}}^T\mathbf{b} \\ &= const + \mathbf{y}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{B}}^T\mathbf{a} + \mathbf{y}^T\widetilde{\mathbf{Y}}\widetilde{\mathbf{A}}^T\mathbf{b}. \end{aligned} \tag{19}$$

Here $Tr(\mathbf{a}\mathbf{y}^T\mathbf{y}\mathbf{b}^T) = Tr(\mathbf{a}\mathbf{b}^T) = const$. Similarly, we have:

$$Tr(\mathbf{C}\mathbf{Y}^T) = Tr(\mathbf{c}\mathbf{y}^T) + Tr(\widetilde{\mathbf{C}}\widetilde{\mathbf{Y}}^T) = const + \mathbf{y}^T\mathbf{c}. \qquad (20)$$

According to Eq.(19) and Eq.(20), we can formulate Eq.(18) as follows:

$$\begin{aligned} &\min_{\mathbf{y}} \; \mathbf{y}^T(2\widetilde{\mathbf{Y}}\widetilde{\mathbf{V}}^T\mathbf{v} + \alpha\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}}^T\mathbf{e} + \alpha\widetilde{\mathbf{Y}}\widetilde{\mathbf{E}}^T\mathbf{f} - \mathbf{g} - \mathbf{h}) \\ &s.t. \;\; \mathbf{y} \in \{0,1\}^{c\times 1}, \sum_{i}^{c} \mathbf{y}_i = 1, \end{aligned} \qquad (21)$$

where $\mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}$ are the i-th column of $\mathbf{E}$,$\mathbf{F}$,$\mathbf{G}$,$\mathbf{H}$ and $\widetilde{\mathbf{E}}$,$\widetilde{\mathbf{F}}$,$\widetilde{\mathbf{G}}$,$\widetilde{\mathbf{H}}$ are the matrices of $\mathbf{E}, \mathbf{F}, \mathbf{G}, \mathbf{H}$ excluding $\mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}$ respectively.

Thus, this subproblem can be easily solved as:

$$\mathbf{y}_i = \begin{cases} 1 & \text{i} = \text{h}(\mathbf{m}) \\ 0 & \text{otherwise}, \end{cases} \qquad (22)$$

where $\mathbf{m} = 2\widetilde{\mathbf{Y}}\widetilde{\mathbf{V}}^T\mathbf{v} + \alpha\widetilde{\mathbf{Y}}\widetilde{\mathbf{F}}^T\mathbf{e} + \alpha\widetilde{\mathbf{Y}}\widetilde{\mathbf{E}}^T\mathbf{f} - \mathbf{g} - \mathbf{h}$, and $\text{h}(\mathbf{m})$ returns the index of the minimum value of $\mathbf{m}$. After $2 \sim 3$ inner iterations, we can obtain a optimal complete $\mathbf{Y}$.

3) Solve $\mathbf{v}$ *when* $\mathbf{U}_a, \mathbf{U}_b, \mathbf{Y}$ *are fixed.* The loss of each sample can be directly computed, thus we can obtain $\mathbf{v}$ by Eq.(13).

Algorithm 1 Self-Paced Cross-Modal Subspace Matching (SCSM)

```
Input: The matrices of unlabeled data $\mathbf{X}_p \in \mathbb{R}^{d_p \times n}, p \in \{a, b\}$;
Output: The projection matrices $\mathbf{U}_p \in \mathbb{R}^{d_p \times c}, p \in \{a, b\}$;
1: Initialize $\mathbf{Y}$ by K-means clustering, and compute the Laplacian matrix of the multimodal graph $\mathbf{L}$, and initialize $t = 0, \mathbf{U}_p, p \in \{a, b\}$ as identity matrix;
2: repeat
3:     Compute $\mathbf{v}^t$ according to Eq.(13);
4:     Compute $\mathbf{U}_p^t, p \in \{a, b\}$ by solving the linear system problems in Eq.(16);
5:     Compute $\mathbf{Y}^t$ according to Eq.(21) and Eq.(22);
6:     Compute $\mathbf{W}^t$ according to Eq.(10);
7:     $t = t + 1$;
8: until Convergence;
```

Algorithm 1 summarizes the alternate minimization procedure to optimize Eq.(14). Firstly, the label information $\mathbf{Y}$ is obtained by K-means clustering on the text modality and the multimodal Laplacian matrix is constructed in Step 1; Then, Step $3 \sim 6$ is an alternate procedure of iteratively updating each of $\mathbf{U}\{a, b\}, \mathbf{v}, \mathbf{Y}$ one by one in a loop.

2.5 Computational Complexity Analysis

Here we try to discuss the time complexity of the proposed SCSM. For each outer iteration, we need $\mathcal{O}(nd_1^2 + nd_2^2)$ to solve $\mathbf{U}_a$ and $\mathbf{U}_b$, $\mathcal{O}(nc^2)$ to solve $\mathbf{Y}$, and $\mathcal{O}(ncd_1 + ncd_2)$ for solving v. Thus, the overall time complexity is $\mathcal{O}(nd^2 + n^2c)$ for SCSM, where c is the dimension of the common subspace and d is the larger original dimensionality for each modality.

3. EXPERIMENT

In this section, we comprehensively compare SCSM and other existing cross-modal approaches on four benchmark datasets, the Pascal VOC [15, 10], the Wikipedia [29], and the LabelMe [32, 26] dataset, which are widely adopted for cross-modal matching tasks.

3.1 Evaluation Criteria & Baseline Methods

There are several widely used evaluation criteria for cross-modal subspace matching algorithms [29, 33, 37]. During the testing phase, multimodal data are mapped into a common subspace via the learned projection matrices for each modality. Then in the subspace, we take one type of modality as a query set to match another type of modality. Like [29, 37, 18], the cosine distance is utilized to measure the similarity of projected features. The simple goal is that given a query from one modality, we can return the top k closest matches in another modality via subspace learning.

The mean average precision (MAP) metric is a classical performance evaluation criterion in the information retrieval circle. The higher MAP indicates the better performance. Besides the MAP, we also exploit the precision-scope curve [30] and precision-recall curve [29] to intensively evaluate the effectiveness of different methods, where the scope K is specified by the number of the top-ranked texts/images presented to users.

With regards to baseline methods, CCA [12], PLS [31] and GMBLM [33] are three unsupervised methods, which merely use the hard pair-wised information to learn a common latent subspace. In contrast, supervised methods such as CDFE [23], GMLDA [33], GMMFA [33], LCFS [37] and recent JFSSL [36] all take the semantic class information into account, and they are also compared here. Despite the improved performance, semantic labels are usually expensive and time-consuming to obtain for supervised methods. To further study the effect of initial clustering result, we further assume the clustering result as pseudo semantic labels, and attain the matching performance of above supervised methods, referred as UCDFE, UGMLDA and UGMMFA.

3.2 Performance on Cross-Modal Matching

3.2.1 *Results on the Pascal VOC dataset*

The Pascal VOC dataset [15] is a commonly used dataset, it consists of 9,963 image-text pairs from 20 categories, including 5,011 training and 4,952 testing image-text pairs. Since some of the images have multiple labels, we select the images with only one label [33, 37]. As a result, we obtain 2808 training and 2841 testing data samples that correspond to 20 categories. The image and text features are 512-dimensional Gist [26] features and 399-dimensional word frequency features, respectively.

We first use the Principal Component Analysis (PCA) to remove the redundancy in features. 95% information energy is preserved by the PCA before evaluation. The MAP scores of the cross-modal retrieval results are shown in Table 1.

From Table 1, we can observe that the proposed SCSM achieves the best performance on both the image-to-text and text-to-image matching tasks. Obviously, supervised methods (i.e., CDFE, GMLDA and GMMFA) outperform their unsupervised versions by 10%, and earn bigger advantages over CCA and PLS. This is because supervised methods rely on semantic labels to reduce the semantic gap of different modalities, but unsupervised methods only use pair-wised information. However, our unsupervised method not only surpasses the unsupervised methods,

Dataset & Methods	Pascal VOC			Wiki			Wiki++			LabelMe		
	Image	Text	Avg	Image	Text	Avg	Image	Text	Avg	Image	Text	Avg
CCA	0.250	0.212	0.231	0.251	0.199	0.225	0.347	0.310	0.329	0.268	0.236	0.252
PLS	0.256	0.241	0.249	0.262	0.174	0.218	0.304	0.329	0.317	0.522	0.435	0.478
GMBLM	0.312	0.232	0.272	0.255	0.204	0.229	0.347	0.318	0.333	0.515	0.466	0.490
UCDFE	0.279	0.209	0.244	0.224	0.184	0.204	0.333	0.301	0.317	0.570	0.594	0.582
UGMMFA	0.298	0.232	0.265	0.269	0.211	0.240	0.340	0.310	0.325	0.512	0.499	0.505
UGMLDA	0.301	0.239	0.270	0.272	0.215	0.244	0.340	0.318	0.325	0.542	0.536	0.539
CDFE	0.306	0.227	0.267	0.260	0.209	0.234	0.397	0.344	0.370	0.685	**0.725**	0.705
GMMFA	0.327	0.259	0.293	0.273	0.219	0.246	0.409	0.362	0.386	**0.719**	0.724	**0.722**
GMLDA	0.324	0.260	0.292	0.273	0.218	0.246	0.409	0.362	0.386	0.716	0.720	0.718
LCFS [36]	0.344	0.267	0.306	0.280	0.214	0.247	0.413	0.384	0.404	-	-	-
JFSSL [36]	**0.361**	**0.280**	**0.320**	**0.306**	**0.228**	**0.267**	**0.428**	**0.396**	**0.412**	-	-	-
SCSM	**0.375**	**0.282**	**0.329**	**0.274**	**0.217**	**0.245**	**0.423**	**0.381**	**0.402**	**0.641**	**0.672**	**0.656**

Table 1: MAP scores of *unsupervised* SCSM and other methods on the Pascal VOC, Wiki, Wiki++ and LabelMe datasets, while CDFE, GMMFA, GMLDA, LCFS and JFSSL are *supervised* methods.

but outperforms several supervised methods, achieving the state-of-the-art performance.

Figure 2 shows the corresponding precision-recall curves and precision-scope curves. The scope, i.e., the top K retrieved items, varies from $K = 50$ to 1000. It can be observed from Figure 2 that SCSM consistently outperforms other unsupervised methods for both Image query vs. Text database and Text query vs. Image database, regardless of precision-scope curves and precision-recall curves. Note that GMBLM only performs inferior to SCSM, due to its consideration of intra-similarity.

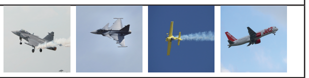
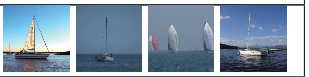
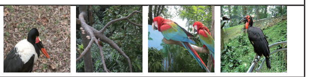
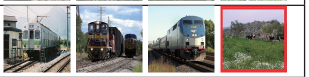
Class	Tags queries	True Image	Matching Images
aeroplane	**aeroplane grass door window wheel**		
boat	**boat rigging water sky wave**		
bird	**bird tree sky**		
train	**train rope railroad pole tree**		

Table 2: Retrieval examples by user tags queries on the Pascal VOC database by the proposed SCSM. For each tags query (second column), the top several retrieved images are shown in the fourth column. The first column represents the ground truth class of each corresponding tags query, and the third column shows the ground truth Image with the queries. (Best viewed in colors.)

3.2.2 Results on the Wiki dataset

The Wiki image-text dataset [29] generated from Wikipedia articles, is a fundamental dataset for cross-modal matching. It is composed of 2,866 image-text pairs from 10 semantic classes, with image features being 128-dimensional SIFT feature and the text feature being 10-dimensional Latent Dirichlet allocation (LDA) feature. We split it into a training set of 1,300 pairs and a testing set of 1,566 pairs in the experiment [37]. We directly carry out the experiment with the provided datasets owing to the low dimensional image and text features. The MAP scores obtained by SCSM and other approaches are shown in Table 1.

From Table 1, we can see that SCSM achieves similar results to GMLDA and GMMFA, and is just inferior to JFSSL. However, it continues to be the best algorithm for unsupervised cross-modal subspace learning. The reason for these results is that the dimension of the features in this database is generally low so that the feature learning hardly takes effect. In the next subsection, we will show our improved performances with the help of better feature representations.

The corresponding precision-recall curves (a) and precision-scope curves (b) are also plotted for both forms of cross-modal retrieval tasks in Figure 3. We observe that, for an image query, SCSM obtains similar and even inferior performance, while for the text query, it performs the best among these unsupervised algorithms.

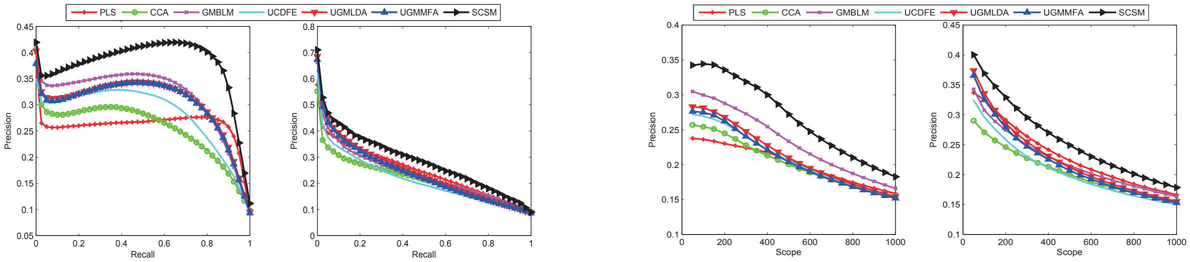
3.2.3 Results on the Wiki++ dataset

The Wiki++ dataset [36] is built on the famous Wiki dataset [29]. They share all but feature representations. 4,096-dimensional Convolutional Neural Network (CNN) features are extracted for images by Caffe[1], and 5,000-dimensional BOW feature vectors are learned based on the basic tf-idf features. We also split the entire dataset into a 1,300 training set and a 1,566 testing set. Specifically, PCA is also performed on the features, and 95% information energy preserved for both views.

From Table 1, it is obvious that the proposed SCSM algorithm achieves the best performance with average MAP 40.2% among unsupervised methods. Even though SCSM performs a little lower than JFSSL, it still can beat other supervised methods such as GMLDA and LCFS. Compared with performances obtained in the Wiki dataset, all methods achieve better results when using CNN visual features in spite of the image query or the text query. This is because the CNN features have proven effective for image feature representation. In the high and sparse dimensional feature setting, SCSM shrinks the distance with the best-performing JFSSL from 2.2% to 1%, offers the great potentials when better features are provided.

SCSM obtains the best performance in terms of precision-

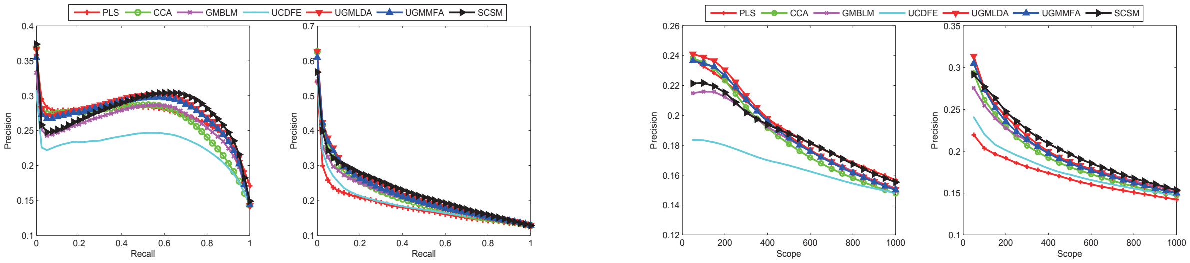
[1] http://caffe.berkeleyvision.org/.



(a) precision-recall curve

(b) precision-scope curve

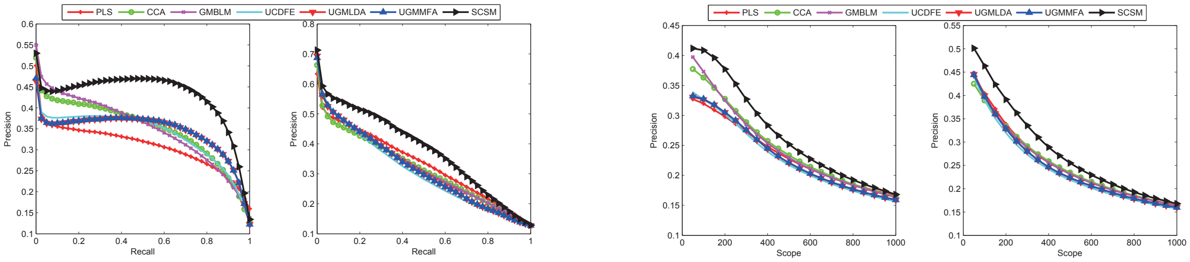
Figure 2: Performance of various *unsupervised* methods on the **Pascal VOC** dataset, based on precision-recall curve(a) for $K = 50$ to 1000 and precision-scope curve(b). The performances for image query are shown in the left sub-figures.



(a) precision-recall curve

(b) precision-scope curve

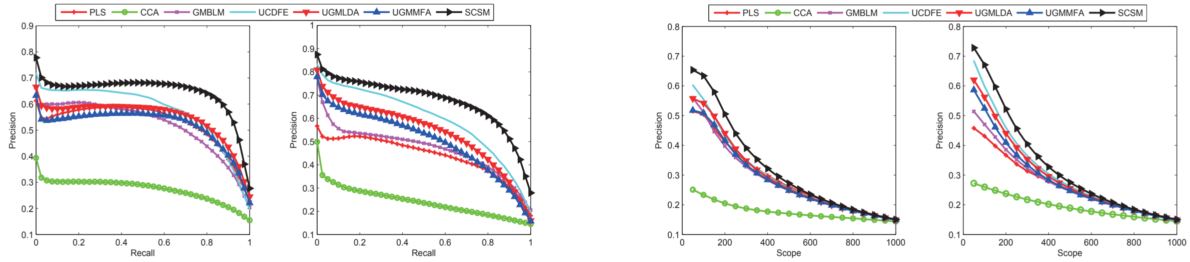
Figure 3: Performance of various *unsupervised* methods on the **Wiki** dataset, based on precision-recall curve(a) for $K = 50$ to 1000 and precision-scope curve(b). The performances for image query are shown in the left sub-figures.



(a) precision-recall curve

(b) Precision-scope curve

Figure 4: Performance of various *unsupervised* methods on the **Wiki++** dataset, based on precision-recall curve(a) for $K = 50$ to 1000 and precision-scope curve(b). The performances for image query are shown in the left sub-figures.



(a) precision-recall curve

(b) precision-scope curve

Figure 5: Performance of various *unsupervised* methods on the **LabelMe** dataset, based on precision-recall curve(a) for $K = 50$ to 1000 and precision-scope curve(b). The performances for image query are shown in the left sub-figures.

recall curves for both image query and text query, as shown in Figure 4(a). From the precision-scope curves in Figure 4(b), SCSM discovers more good matches in the top K documents than its several counterparts.

To further study the difference between SCSM and other methods (here we ignore the unsupervised versions of supervised methods due to the space limit) in Figure 6. From the figure, we can find that SCSM takes the first space for 5 classes, all these classes are easily distinguished. For the rest 5 classes, SCSM is a bit lower than supervised methods but higher than CCA, PLS and GMBLM. It can be concluded that SCSM can achieve a comprehensively better performance among unsupervised methods.

Methods	Matching Images
SCSM	
GMLDA	
CCA	
GMBLM	

Table 3: Retrieval examples by text query “In 1994, Angus and Malcolm invited Rudd to several *jam sessions* ...” on the Wiki database by the proposed method. Red border indicates an incorrect matching result. (Best viewed in colors.)

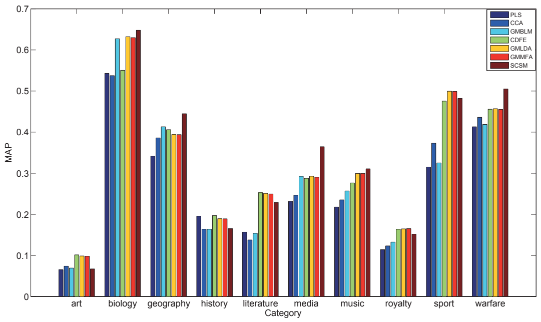


Figure 6: MAP of different methods on the Wiki++ dataset with respect to each category. (Best viewed in colors.)

3.2.4 *Results on the LabelMe dataset*

The LabelMe Outdoor dataset [32, 18] consists of 2,688 fully annotated outdoor images from 8 categories, i.e., “coast”, “forest”, “highway”, “inside city”, “mountain”, “open country” and “tall building”. For the text modality, we generate the object account vector via the LabelMe[2] toolbox. We randomly select 200 samples from each category for training, resulting in 1,600 image-text pairs for the training set and the remaining 1,086[3] image-text pairs for the testing set. There are total 789 unique words with frequencies varying from one time to more than 2000 times. We select the word frequencies that are more than one time. As a result, the image and text features are 512-dimensional Gist features and 470-dimensional word frequency features, respectively.

From Table 1, SCSM obtains the best average score 65.6% among the unsupervised methods, however, it performs worse than some supervised methods. It may be because the word frequency features contain massive noises, such as misspelling and duplicated words, make it hard for unsupervised methods to do feature learning without semantic labels. However, the advanced performances from the Wiki to the Wiki++ database indicate that SCSM can probably approach the supervised methods with deep features for the LabelMe database.

Regarding the precision-recall and precision-scope curves, SCSM still obtains the best performances in both image query and text query tasks. The gap between SCSM and second-best performing UCDFE is apparent, thus we can conclude that SCSM is distinguished, even other algorithms share the same initial pseudo semantic labels.

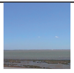
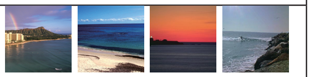
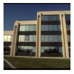
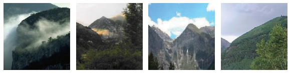
Class	Tags queries	True Image	Matching Images
coast	**sky sea water sand beach**		
insidecity	**building staircase window flag balcony**		
mountain	**clouds mountain land river water island**		
street	**tree sidewalk cars side road streetlight**		

Table 4: Retrieval examples by tags queries on the LabelMe database by the proposed method. For each tags query (second column), the top several retrieved images are shown in the fourth column. The first column represents the ground truth class of each corresponding tags query, and the third column shows the ground truth Image with the queries. (Best viewed in colors.)

Image queries	SCSM	GMLDA
	“aeroplane tree grass insect navy”, “aeroplane leaves tail building sky light”, “sky aeroplane cable”, “aeroplane sky branch”	“sky aeroplane cable”, “aeroplane letter wing moon star door tail sky”, “bicyclehandle wheel clothes pedal leaves ground stairstep”, “aeroplane radiotower cart cone”
	“car tree road sky stone grass”, “pole wire sign trafficlight road fence”, “car ground road hill road”, “car road sky grass tree”	“house tree sky grass curb street car tag”, “car street grass tree house sky”, “car decoration grill carlight”, “bus debris trash house house”
	“foot leg clothes had tabletop”, “rock water waterfall person ground”, “person arm hand foot clothes”, “person clothes water shoe glasses wall”	“sign building trash wall person”, “person ribbon clothes slide fence”, “clothes ground junk”, “frame clothes watch bracelet person”

Table 5: Retrieval examples by image queries on the Pascal VOC database. For each image query (second column), the top several retrieved tags are shown in the second column for the proposed **unsupervised** SCSM and third column for **supervised** GMLDA.

3.3 Convergence & Parameter Analysis

It is obvious to prove that the employed alternative minimization strategy can converge to a local optimum. However, under the self-paced framework, our learning algorithm is hard to guarantee the global convergence. On the other hand, the experimental results in Figure 8 on all four datasets can demonstrate the objective, $\sum_{p\in\{a,b\}}||(\mathbf{U}_p^T\mathbf{X}_p-\mathbf{Y})diag(\mathbf{v})||_F^2 + \alpha Tr(\mathbf{U}_p^T\mathbf{X}_p\mathbf{L}_{pq}\mathbf{X}_q^T\mathbf{U}_q)$, decreases as the iteration number increases.

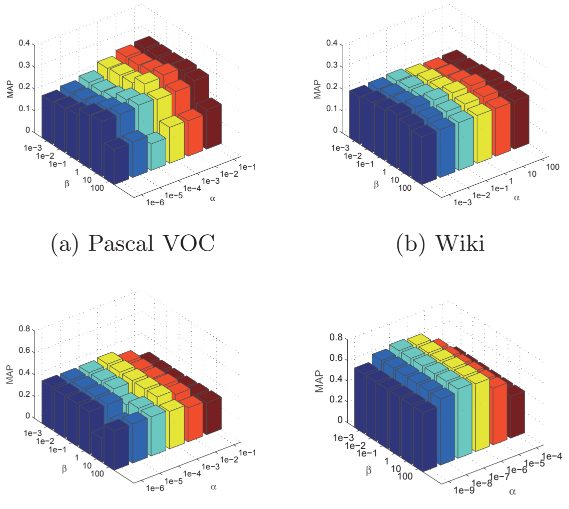
With regards to the parameters α and β in SCSM, we conduct an experiment to study the influence of different

[2] http://labelme.csail.mit.edu/Release3.0/browserTools/php/matlabtool_box.php.

[3] Two instances are dropped due to its missing tags (refrred as text).

Image queries	SCSM	GMLDA
	“water sea building sky sand beach mountain rainbow”, “sand beach water sea sky”, “sky mountain ocean water water sea sand beach trees”, “water sea sky waves”	“sky mountain field river water trees”, “sea water sky sand beach building tree”, “sky buildings sea water”, “sky buildings sea water”
	“sky field road tree shrub”, “sky desert ground trees hill”, “sky snowy mountain hill buildings”, “sky mountain field grass cow”	“sky path shrub trees stone”, “sky mountain field river water trees”, “sky mountain field desert shrub plant”, “sky desert ground trees hill”
	“sky skyscraper building”, “building trees sidewalk road bus”, “sky buildings skyscraper”, “sky skyscraper building”	“sky building wall step person walking”, “sky building car box plants”, “sky skyscraper buildings car side sidewalk”, “city river water building skyscraper dock”

Table 6: Retrieval examples by image queries on the LabelMe database. For each image query (second column), the top several retrieved tags are shown in the second column for the proposed **unsupervised** SCSM and third column for **supervised** GMLDA.



(a) Pascal VOC

(b) Wiki

(c) Wiki++

(d) LabelMe

Figure 7: Performance variation for the average MAP with respect to α and β with γ fixed as 1.

choices. Note that in our experiment, γ is set as 1 while α varies with the training data size n and feature length d, and β varies from $[1e^{-3}, 1e^{-2}, \cdots, 1e^{2}]$. The analysis on parameter sensitivity shows that SCSM is very robust to model parameters which can achieve stable and superior performance under a wide range of parameter values.

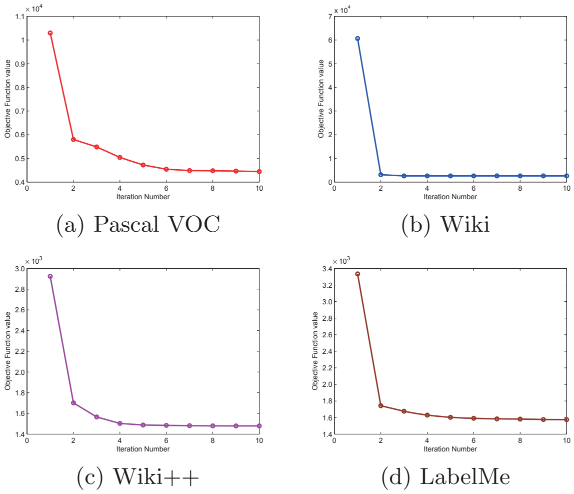
For fair comparisons, we tune the parameter of subspace dimensionality for subspace based cross-modal approaches, and adopt the best-performing dimensionality for comparison. For LCFS, JFSSL and our SCSM, the subspace dimensionality is fixed as the number of semantic classes (e.g., $c = 20$ for the Pascal VOC dataset). Besides, the width parameter σ for Gaussian kernel in Eq.(7) is fixed as 1 for following experiments.

3.4 Discussion

Experimental results show that, our SCSM not only surpasses the unsupervised methods (e.g., CCA, GMBLM), but also outperforms some supervised methods (e.g., CDFE, GMLDA, and LCFS), in the Pascal VOC and Wiki++ datasets. As an unsupervised method, the reasons for the better performance of our SCSM may lie in twofold.

First, images often contain several semantic concepts, potentially belonging to several semantic groups, their labels may be inaccurate for some challenging datasets. Hence the grouping results may be a complementary for the missing label information when feature representations are enough powerful. Second, the pseudo group label obtained by canonical clustering methods is inaccurate. Since clustering is a non-convex problem, we introduce self-paced learning to avoid the local minima and refine the grouping results. Besides, the projection matrices are computed by two regression-like analyses on the pseudo label, and the multimodal graph preserving the inter- and intra- similarities is also constructed.

Moreover, we observe that all methods, especially SCSM, perform better in the Wiki++ dataset. Because the deep features in the Wiki++ dataset are high-dimensional and discriminative, SCSM can more effectively learn features, and estimate the latent semantic group more accurately. However, our method in the Wiki and LabelMe dataset is inferior to the supervised state-of-the-art method proposed by [36], much effort still needs to be taken to improve the performance when the features contain massive noises.



(a) Pascal VOC

(b) Wiki

(c) Wiki++

(d) LabelMe

Figure 8: Convergence curve of the proposed SCSM.

4. CONCLUSION

In this paper, we have proposed a novel *unsupervised* method for cross-modal subspace matching. We have introduced hard pair-wised constraints and soft semantic group constraints for multi-modal data, which are potentially effective for unsupervised learning. A joint feature learning and data grouping formulation has been accordingly developed. For an accurate estimation of the semantic group, self-paced learning has been incorporated into the non-convex loss. Moreover, multimodal graph is included to preserve the inter- and intra- modality similarity. To minimize this joint learning problem, we have presented an alternating minimization solution. Experimental results on four multimodal databases demonstrate that the effectiveness of the proposed method for unsupervised multimodal data. For future work, the proposed method will be extended for multi-label cross-modal matching problem.

Acknowledgments

This work was supported by the National Basic Research Program of China (Grant No. 2012CB316300), the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) (Grant No. 2015190) and the National Natural Science Foundation of China (Grant No. 61473289). Jian Liang and Zhihang Li contributed equally to this work and should be considered co-first authors.

5. REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *ICML*, pages 1247–1255, 2013.

[2] F. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In *NIPS*, pages 49–56, 2008.

[3] S. Basu and J. Christensen. Teaching classification boundaries to humans. In *AAAI*, pages 109–115, 2013.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, pages 41–48. ACM, 2009.

[5] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR*, pages 127–134, 2003.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

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

[8] P. Dhillon, D. P. Foster, and L. H. Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207, 2011.

[9] P. S. Dhillon, J. Rodu, D. P. Foster, and L. H. Ungar. Two step cca: A new spectral method for estimating vector models of words. In *ICML*, pages 1551–1558, 2012.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[11] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2014.

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

[13] R. He, M. Zhang, L. Wang, Y. Ji, and Q. Yin. Cross-modal subspace learning via pairwise constraints. *IEEE TIP*, 24(12):5543–5556, 2015.

[14] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, pages 153–160, 2003.

[15] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *IEEE TPAMI*, 34(6):1145–1158, 2012.

[16] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, pages 2407–2414, 2011.

[17] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *MM*, pages 547–556, 2014.

[18] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE TMM*, 17(3):370–381, 2015.

[19] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, pages 3276–3284, 2015.

[20] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, pages 1189–1197, 2010.

[21] A. Li, S. Shan, X. Chen, B. Ma, S. Yan, and W. Gao. Cross-pose face recognition by canonical correlation analysis. *arXiv preprint: 1507.08076*, 2015.

[22] J. Liang, R. He, Z. Sun, and T. Tan. Group-invariant cross-modal subspace learning. In *IJCAI*, 2016.

[23] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, pages 13–26. 2006.

[24] J. Masci, M. M. Bronstein, A. Bronstein, and J. Schmidhuber. Multimodal similarity-preserving hashing. *IEEE TPAMI*, 36(4):824–830, 2014.

[25] D. Meng and Q. Zhao. What objective does self-paced learning indeed optimize? *arXiv preprint:1511.06049*, 2015.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[27] B. Ozdemir and L. S. Davis. A probabilistic framework for multimodal retrieval using integrative indian buffet process. In *NIPS*, pages 2384–2392, 2014.

[28] D. Putthividhy, H. T. Attias, and S. S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, pages 3408–3415, 2010.

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

[30] N. Rasiwasia, P. J. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *IEEE TMM*, 9(5):923–938, 2007.

[31] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection*, pages 34–51. Springer, 2006.

[32] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[33] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, pages 2160–2167, 2012.

[34] F. Shen, C. Shen, W. Liu, and H. Tao Shen. Supervised discrete hashing. In *CVPR*, pages 37–45, 2015.

[35] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2222–2230, 2012.

[36] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE TPAMI*, 2015. doi:10.1109/TPAMI.2015.2505311.

[37] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *ICCV*, pages 2088–2095, 2013.

[38] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE TCYB*, 2016. doi:10.1109/TCYB.2016.2519449.

[39] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang. Quantized correlation hashing for fast cross-modal search. In *IJCAI*, pages 3946–3952, 2015.

[40] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *SIGIR*, pages 395–404, 2014.

[41] T. Zhang and J. Wang. Collaborative quantization for cross-modal similarity search. In *CVPR*, 2016.

[42] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, pages 3196–3202, 2015.

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

[44] Y. T. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. M. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.