Learning Coupled Feature Spaces for Cross-modal Matching

Kaiye Wang, Ran He, Wei Wang, Liang Wang, Tieniu Tan
Center for Research on Intelligent Perception and Computing, National Lab of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{kaiye.wang, rhe, wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Cross-modal matching has recently drawn much attention due to the widespread existence of multimodal data. It aims to match data from different modalities, and generally involves two basic problems: the measure of relevance and coupled feature selection. Most previous works mainly focus on solving the first problem. In this paper, we propose a novel coupled linear regression framework to deal with both problems. Our method learns two projection matrices to map multimodal data into a common feature space, in which cross-modal data matching can be performed. And in the learning procedure, the ℓ_{21}-norm penalties are imposed on the two projection matrices separately, which leads to select relevant and discriminative features from coupled feature spaces simultaneously. A trace norm is further imposed on the projected data as a low-rank constraint, which enhances the relevance of different modal data with connections. We also present an iterative algorithm based on half-quadratic minimization to solve the proposed regularized linear regression problem. The experimental results on two challenging cross-modal datasets demonstrate that the proposed method outperforms the state-of-the-art approaches.

1. Introduction

This paper focuses on the cross-modal matching problem, which has been widely studied in recent years. The task of cross-modal matching is to predict whether a pair of data points from two different modalities represent the same underlying content or object. The cross-modal problem has been existing in many fields. Take multimedia retrieval for example, one often seeks to find the picture (or video) that best illustrates a given text, or find the text that best describes a given picture (or video).

As shown in the next section, there have been some methods proposed for solving the cross-modal problems. Most of them just focus on learning a common latent subspace to make all data comparable. However, another important problem, how to simultaneously select relevant and discriminative features from two different feature spaces, is usually ignored. Here we call this problem "coupled feature selection". Although various feature selection methods [26] have been developed for the single modality data analysis, they are not extended to the case of multi-modality data.

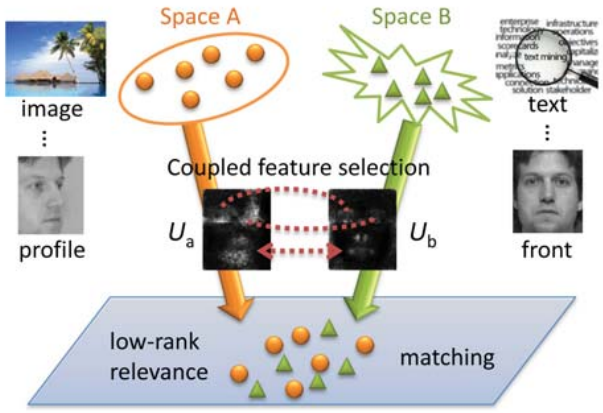

Figure 1. The overview of the proposed method. U_A and U_B are projection matrices learned using our method on space A and B. ℓ_{21}-norm and trace norm are used for coupled feature selection and low-rank relevance measure respectively.

Recently, ℓ_{21}-norm has been proved to be a powerful tool for the feature selection problem [5, 8, 15], and trace norm [1, 3, 4, 6] is used to encode the correlation of the design matrix or prior knowledge by enforcing a low-rank solution. Motivated by these recent advances, this paper proposes a novel regularization framework (as shown in Figure 1) for the cross-modal matching problem, by combining common subspace learning and coupled feature selection. First, inspired by the potential relationship between Canonical Correlation Analysis (CCA) and linear least squares [23], coupled linear regression is used to project data from different modalities into a common subspace that is defined by label information. In the same time, ℓ_{21}-norm is used to select the relevant and discriminative features from coupled

modalities, and the trace norm regularization enforces the relevance of the projected data with potentially connections. Second, based on the alternative formulation for the trace norm [4] and the half-quadratic analysis for ℓ_{21}-norm [8], we develop an iterative algorithm to solve the proposed regularization problem. Finally, the proposed method is applied to text-image retrieval and experimental results on two public datasets show that the proposed method outperforms previous approaches.

Main contributions of our work can be summarized as follows:

1) A novel regularization framework is proposed for the cross-modal matching problem. It unifies coupled linear regressions, ℓ_{21}-norm and trace norm into a generic minimization formulation so that subspace learning and coupled feature selection can be performed simultaneously.

2) An iterative algorithm is presented to efficiently solve such kind of complex minimization problems. In each iteration, the minimization problem is simplified to two independent linear system problems. And we prove the convergence of the proposed optimization method.

3) The proposed framework provides a new way to effectively deal with a challenging cross-modal problem, i.e., text-image retrieval. Experimental results on two public cross-modal datasets have shown that our proposed framework outperforms several relevant state-of-the-art approaches.

The remainder of the paper is organized as follows. In Section 2, we overview related work on the cross-modal problem. Section 3 describes our proposed regularized linear regression framework for cross-modal matching, along with an iterative algorithm to solve this problem. In Section 4, we report experimental results on two cross-modal datasets. Finally, we conclude the paper in Section 5.

2. Related Work

Since the cross-modal matching is considered as an important problem in some real applications, various approaches have been proposed to deal with this problem, such as Canonical Correlation Analysis (CCA) [7], Partial Least Squares (PLS) [20], and Bilinear Model (BLM) [22, 24]. Specifically, CCA is probably the most popular one due to its wide-spread use in cross-media retrieval[7, 19], cross-lingual retrieval [25] and some vision problems [13, 14]. Rasiwasia et al. [19] apply CCA to the cross-modal mutimedia retrieval. Based on the hypothesis that there is a benefit to explicitly model correlations between two modalities, CCA is used to learn a common subspace by maximizing the correlation between the two modalities. Then, a semantic space is learned to measure the similarity of different modal features. Li et al. [14] apply CCA to face recognition based on non-corresponding region matching. They use CCA to learn a common space in which the possibility of whether two non-corresponding face regions belong to the same face can be measured. Recently, Partial Least Squares (PLS) [20] is also used for the cross-modal matching problem. To perform multi-modal face recognition, such as front vs. profile, photos vs. sketches, and high-resolution photos vs. low-resolution photos, Sharma and Jacobs [21] use PLS to linearly map images in different modalities to a common linear subspace in which they are highly correlated. Chen et al. [2] apply PLS to the cross-modal document retrieval. They use PLS to switch the image features into the text space, then learn a semantic space for the measure of similarity between two different modalities . In [24], Tenenbaum and Freeman propose a bilinear model (BLM) to derive a common space for cross-modal face recognition, and BLM is also used for text-image retrieval in [22].

Besides CCA, PLS and BLM, there are some other methods for the cross-modal problem. Lei and Li [12] propose coupled spectral regression to learn two associated projections, which project heterogeneous data into a common space respectively in which classification is performed. Quadrianto and Lampert [17] propose a metric learning approach for cross-modal matching, which considers both matching and non-matching samples. Huang et al. [10] propose a unified framework extended from graph embedding and design an algorithm for face recognition across poses or resolutions. Recently, Sharma et al. [22] extend Linear Discriminant Analysis (LDA) and Marginal Fisher Analysis (MFA) to their multiview counterpart, i.e., Generalized Multiview LDA (GMLDA) and Generalized Multiview MFA (GMMFA), and apply them to deal with the cross-media retrieval problem.

All above methods can be categorized into two classes: one is to learn a common latent space in which both modalities are projected, and the other is to map data of one modality into the space of another one. They all focus on measurement of relevance, however, ignore another important issue, i.e., coupled feature selection. Since the dimensionality of real world data is often high, there are naturally many redundant and irrelevant features. Hence, how to simultaneously select the relevant and discriminative features for different modalities of data is very important. Accordingly, we aim to jointly perform common subspace learning and coupled feature selection. To achieve this goal, we propose a generic minimization formulation by coupled linear regressions, ℓ_{21}-norm and trace norm, which will be detailed in the next section.

3. Learning Coupled Feature Spaces

In this section, we present a novel framework for the cross-modal matching problem, which can be formulated as a minimization problem. Then, an iterative algorithm based on half-quadratic optimization is given to solve this minimization problem. We begin with a brief introduction

to some notations.

Notations. Let $\mathbf{M} \in \mathbb{R}^{n\times m}$. For matrix $\mathbf{M}$, its i-th row, j-th column are denoted by $\mathbf{M}^{(i)}$, $\mathbf{M}_j$ respectively. The Frobenius norm of the matrix $\mathbf{M}$ is defined as

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{n} \left\|\mathbf{M}^{(i)}\right\|_2^2} \tag{1}$$

$\|\mathbf{M}\|_{2,1}$ is the sum of the ℓ_2-norm of the rows of $\mathbf{M}$:

$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n} \left\|\mathbf{M}^{(i)}\right\|_2 \tag{2}$$

$\|\mathbf{M}\|_*$ is the trace norm, i.e., the sum of the singular values of the matrix $\mathbf{M}$, defined as follows

$$\|\mathbf{M}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i \tag{3}$$

where σ_i denotes the i-th singular value of $\mathbf{M}$. For $\mathbf{M} \in \mathbb{R}^{m\times m}$, $diag(\mathbf{M}) \in \mathbb{R}^m$ is the diagonal of the matrix $\mathbf{M}$, while for $\mathbf{u} \in \mathbb{R}^m$, $Diag(\mathbf{u}) \in \mathbb{R}^{m\times m}$ is the diagonal matrix whose diagonal elements are u_i.

3.1. Problem formulation

Suppose that we have a collection of data from two different modalities, each pair $\{\mathbf{x}_i^a, \mathbf{x}_i^b\}$ represents the same underlying content. For example, user tags (or textual descriptions) and image features indicate the same objects or content contained in the image. Given a query from one modality, the goal of the cross-modal matching is to return the closest match in another modality.

As shown in Figure 1, the cross-modal matching generally involves two problems: 1) The first problem is how to measure the relevance of data from different modalities. 2) The second one is how to select the relevant and discriminative features from the coupled feature spaces, simultaneously. Previous methods mainly focus on the first problem, such as CCA or PLS. They project data from different modalities into a latent space, in which the possibility of whether two different modal data represent the same semantic concept can be measured. However, the second problem is usually ignored. Compared to dimensionality reduction or feature selection methods performed on the two feature spaces separately, coupled feature selection is more likely to find the most relevant features. Based on this consideration, we propose that the feature selection procedure should be performed on coupled feature spaces simultaneously for better matching.

Given data from two different modalities: $\mathbf{X}_a = [\mathbf{x}_1^a, \mathbf{x}_2^a, ..., \mathbf{x}_n^a] \in \mathbb{R}^{d1\times n}$, $\mathbf{X}_b = [\mathbf{x}_1^b, \mathbf{x}_2^b ..., \mathbf{x}_n^b] \in \mathbb{R}^{d2\times n}$, each modality has n samples embedded in different dimensional spaces ($d1$ and $d2$), and each pair $\{\mathbf{x}_i^a, \mathbf{x}_i^b\}$ represents the same underlying content and belongs to the same class. Let $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T \in \mathbb{R}^{n\times c}$ be the class label matrix, where c is the number of classes. Our model aims to learn two projection matrices to map the data of the coupled spaces into the common space defined by class labels. In the same time, we perform ℓ_{21}-norm on the projection matrices for coupled feature selection, and impose a low-rank constraint, defined by the trace norm, on the projected data. That is, we have a generic minimization problem in the following form,

$$\begin{aligned}&\min_{\mathbf{U}_a,\mathbf{U}_b} \frac{1}{2}(\left\|\mathbf{X}_a^T\mathbf{U}_a - \mathbf{Y}\right\|_F^2 + \left\|\mathbf{X}_b^T\mathbf{U}_b - \mathbf{Y}\right\|_F^2) \\ &+\lambda_1(\|\mathbf{U}_a\|_{21} + \|\mathbf{U}_b\|_{21}) + \lambda_2\left\|[\mathbf{X}_a^T\mathbf{U}_a \;\; \mathbf{X}_b^T\mathbf{U}_b]\right\|_*\end{aligned} \tag{4}$$

where $\mathbf{U}_a$ and $\mathbf{U}_b$ are the projection matrices for the coupled spaces respectively. The first term is coupled linear regression, which is used to learn two projection matrices for mapping different modal data into a common space. The second term contains two ℓ_{21}-norms that play a role of feature selection on two feature spaces simultaneously. And the trace norm is to enforce the relevance of projected data with connections.

3.2. An iterative solution

It is very complicated to directly minimize the above objective function involving ℓ_{21}-norm and trace norm. Here, an iterative algorithm based on the half-quadratic minimization [8, 9] is proposed to solve this problem. Toward this end, we first need to introduce a variational formulation for the trace norm [4]:

Lemma 1. *Let* $\mathbf{M} \in \mathbb{R}^{n\times m}$. *The trace norm of* $\mathbf{M}$ *is equal to:*

$$\|\mathbf{M}\|_* = \frac{1}{2}\inf_{\mathbf{S}\succcurlyeq 0} \; tr(\mathbf{M}^T\mathbf{S}^{-1}\mathbf{M}) + tr(\mathbf{S}) \tag{5}$$

and the infimum is attained for $\mathbf{S} = \left(\mathbf{M}\mathbf{M}^T\right)^{1/2}$.

Using this lemma, we can reformulate (4) as

$$\begin{aligned}&\min_{\mathbf{U}_a,\mathbf{U}_b} \min_{\mathbf{S}\succcurlyeq 0} \; f(\mathbf{U}_a, \mathbf{U}_b) + \lambda_1(\|\mathbf{U}_a\|_{21} + \|\mathbf{U}_b\|_{21}) \\ &+\frac{\lambda_2}{2}tr([\mathbf{X}_a^T\mathbf{U}_a \;\; \mathbf{X}_b^T\mathbf{U}_b]^T\mathbf{S}^{-1}[\mathbf{X}_a^T\mathbf{U}_a \;\; \mathbf{X}_b^T\mathbf{U}_b]) + \frac{\lambda_2}{2}tr(\mathbf{S})\end{aligned} \tag{6}$$

where $f(\mathbf{U}_a, \mathbf{U}_b)$ denotes the first term of the objective function. Since

$$tr(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}) = tr(\mathbf{A}) + tr(\mathbf{D}) \tag{7}$$

The third term in (6) can be rewritten in the following form:

$$\begin{aligned}&tr([\mathbf{X}_a^T\mathbf{U}_a \;\; \mathbf{X}_b^T\mathbf{U}_b]^T\mathbf{S}^{-1}[\mathbf{X}_a^T\mathbf{U}_a \;\; \mathbf{X}_b^T\mathbf{U}_b]) \\ &\Rightarrow tr(\mathbf{U}_a^T\mathbf{X}_a\mathbf{S}^{-1}\mathbf{X}_a^T\mathbf{U}_a) + tr(\mathbf{U}_b^T\mathbf{X}_b\mathbf{S}^{-1}\mathbf{X}_b^T\mathbf{U}_b)\end{aligned} \tag{8}$$

In order to alternately minimize the objective function over $\mathbf{U}_a, \mathbf{U}_b$ and $\mathbf{S}$, we need to add a term $\frac{\lambda_2 \mu}{2} tr(\mathbf{S}^{-1})$ as in [4] . Otherwise, the infimum over $\mathbf{S}$ could be attained at a non-invertible $\mathbf{S}$, leading to a non-convergent algorithm. The infimum over $\mathbf{S}$ is then attained for

$$\mathbf{S} = (\mathbf{X}_a^T \mathbf{U}_a \mathbf{U}_a^T \mathbf{X}_a + \mathbf{X}_b^T \mathbf{U}_b \mathbf{U}_b^T \mathbf{X}_b + \mu \mathbf{I})^{1/2} \quad (9)$$

If we define $\phi(x) = \sqrt{x^2 + \varepsilon}$, we can replace $\|\mathbf{U}_a\|_{21}$ and $\|\mathbf{U}_b\|_{21}$ with $\sum_i^{d1} \phi(\|\mathbf{u}_a^i\|_2)$ and $\sum_i^{d2} \phi(\|\mathbf{u}_b^i\|_2)$ respectively, according to the analysis for the ℓ_{21}-norm in [8]. And ε is a smoothing term, which is usually set to be a small value. It can be proved that $\phi(x) = \sqrt{x^2 + \varepsilon}$ satisfies all conditions as follows.

$$\begin{aligned} &x \to \phi(x) \text{ is convex on } R, \\ &x \to \phi(\sqrt{x}) \text{ is concave on } R_+, \\ &\phi(x) = \phi(-x), \forall x \in R, \\ &\phi(x) \text{ is } C^1 \text{ on } R, \\ &\phi^{''}(0^+) > 0, \ \lim_{x \to \infty} \phi(x)/x^2 = 0. \end{aligned} \quad (10)$$

Then, we can optimize $\phi(\cdot)$ in a half-quadratic way [16] according to the following Lemma 2 [8].

Lemma 2. *Let $\phi(\cdot)$ be a function satisfying all conditions in (10), for a fixed $\|\mathbf{u}^i\|_2$, there exists a dual potential function $\varphi(\cdot)$, such that*

$$\phi(\|\mathbf{u}^i\|_2) = \inf_{p \in R} \left\{ p \|\mathbf{u}^i\|_2^2 + \varphi(p) \right\} \quad (11)$$

where p is determined by the minimizer function $\varphi(\cdot)$ with respect to $\phi(\cdot)$.

According to Lemma 2, the objective function (6) can be reformulated as follows.

$$\begin{aligned} \min_{\mathbf{U}_a, \mathbf{U}_b} \min_{\mathbf{S} \succcurlyeq 0} & f(\mathbf{U}_a, \mathbf{U}_b) + \lambda_1 (tr(\mathbf{U}_a^T \mathbf{P} \mathbf{U}_a) + tr(\mathbf{U}_b^T \mathbf{Q} \mathbf{U}_b)) \\ & + \frac{\lambda_2}{2} (tr(\mathbf{U}_a^T \mathbf{X}_a \mathbf{S}^{-1} \mathbf{X}_a^T \mathbf{U}_a) + tr(\mathbf{U}_b^T \mathbf{X}_b \mathbf{S}^{-1} \mathbf{X}_b^T \mathbf{U}_b)) \\ & + \frac{\lambda_2}{2} tr(\mathbf{S})) \end{aligned} \quad (12)$$

Given $\mathbf{S}$, optimizing the objective function (12) over $\mathbf{U}_a$ and $\mathbf{U}_b$ respectively is equal to optimizing the following two problems, according to the half-quadratic analysis for ℓ_{21}-norm [8].

$$\begin{cases} \min_{\mathbf{U}_a} \frac{1}{2} \|\mathbf{X}_a^T \mathbf{U}_a - \mathbf{Y}\|_F^2 + \lambda_1 tr(\mathbf{U}_a^T \mathbf{P} \mathbf{U}_a) \\ \qquad + \frac{\lambda_2}{2} tr(\mathbf{U}_a^T \mathbf{X}_a \mathbf{S}^{-1} \mathbf{X}_a^T \mathbf{U}_a) \\ \min_{\mathbf{U}_b} \frac{1}{2} \|\mathbf{X}_b^T \mathbf{U}_b - \mathbf{Y}\|_F^2 + \lambda_1 tr(\mathbf{U}_b^T \mathbf{Q} \mathbf{U}_b) \\ \qquad + \frac{\lambda_2}{2} tr(\mathbf{U}_b^T \mathbf{X}_b \mathbf{S}^{-1} \mathbf{X}_b^T \mathbf{U}_b) \end{cases} \quad (13)$$

where $\mathbf{P} = Diag(\mathbf{p})$ and $\mathbf{Q} = Diag(\mathbf{q})$. And $\mathbf{p}$ and $\mathbf{q}$ are auxiliary vectors of the two ℓ_{21}-norms, respectively. The elements of $\mathbf{p}$ and $\mathbf{q}$ are computed respectively as follows.

$$\begin{cases} p_i = \frac{1}{2\sqrt{\|\mathbf{u}_a^i\|_2^2 + \varepsilon}} \\ q_i = \frac{1}{2\sqrt{\|\mathbf{u}_b^i\|_2^2 + \varepsilon}} \end{cases} \quad (14)$$

where ε is a smoothing term, which is usually set to be a small constant value.

Then, the optimal solution of (13) can be computed via solving the following two linear system problems.

$$\begin{cases} (\mathbf{X}_a \mathbf{X}_a^T + \lambda_1 \mathbf{P} + \lambda_2 \mathbf{X}_a \mathbf{S}^{-1} \mathbf{X}_a^T) \mathbf{U}_a = \mathbf{X}_a \mathbf{Y} \\ (\mathbf{X}_b \mathbf{X}_b^T + \lambda_1 \mathbf{Q} + \lambda_2 \mathbf{X}_b \mathbf{S}^{-1} \mathbf{X}_b^T) \mathbf{U}_b = \mathbf{X}_b \mathbf{Y} \end{cases} \quad (15)$$

Algorithm 1: Iterative Algorithm for Learning Coupled Feature Spaces (LCFS)

Input: $\mathbf{X}_a \in \mathbb{R}^{d1 \times n}$, $\mathbf{X}_b \in \mathbb{R}^{d2 \times n}$ and $\mathbf{Y} \in \mathbb{R}^{n \times c}$

Output: $\mathbf{U}_a \in \mathbb{R}^{d1 \times c}$ and $\mathbf{U}_b \in \mathbb{R}^{d2 \times c}$

Set $t = 0$. Initialize $\mathbf{U}_a$ and $\mathbf{U}_b$ as zero matrix.

repeat

1. Compute $\mathbf{V} Diag(s_k) \mathbf{V}^T$ as the eigenvalue decomposition of $(\mathbf{X}_a^T \mathbf{U}_a \mathbf{U}_a^T \mathbf{X}_a + \mathbf{X}_b^T \mathbf{U}_b \mathbf{U}_b^T \mathbf{X}_b)$.
2. Set $\mathbf{S}^{-1} = \mathbf{V} Diag(1/\sqrt{s_k + \mu}) \mathbf{V}^T$.
3. Compute p_i^t and q_i^t according to (14)
4. Compute $\mathbf{U}_a^t$ and $\mathbf{U}_b^t$ by solving the two linear system problems in (15).
5. $t = t + 1$

until Converges

Algorithm 1 summarizes the alternate minimization procedure to optimize (4). Step 1 and Step 2 correspond to the trace norm, which is expected to reinforce the relevance of projected data of different modalities with connections. In Step 3, we compute the auxiliary vectors $\mathbf{p}$ and $\mathbf{q}$ that correspond to the two ℓ_{21}-norms and play an important role in coupled feature selection. In Step 4, we find the optimal solution $\mathbf{U}_a$ and $\mathbf{U}_b$.

3.3. Convergence and complexity

According to half-quadratic minimization, the objective function is minimized in each step, and will decrease step by step until it achieves an optimal solution.

Proposition 1. Assume f be the objective function, and let $F^{\mathrm{t}} \doteq f(U_I^t, S^t, p^t, q^t)$, then $\{F^t\}_{t=1,\ldots}$ generated by Algorithm 1 converges.

Query	PCA+PLS	PCA+BLM	PCA+CCA	PCA+GMMFA	GMMFA	PCA+GMLDA	LCFS
Image	0.2757	0.2667	0.2655	0.3090	0.2253	0.2418	**0.3438**
Text	0.1997	0.2408	0.2215	0.2308	0.1695	0.2038	**0.2674**
Average	0.2377	0.2538	0.2435	0.2699	0.1974	0.2228	**0.3056**

Table 1. Comparison of MAP (Mean Average Precision) performance of different methods on the Pascal VOC dataset. And PCA is performed on the original features to remove redundant features.

Proof. According to the definition of F^t, we have the following equation:

$$\begin{aligned} F^t - F^{t-1} &= \{f(U_I^t, S^t, p^t, q^t) - f(U_I^{t-1}, S^t, p^t, q^t)\} \\ &+\{f(U_I^{t-1}, S^t, p^t, q^t) - f(U_I^{t-1}, S^t, p^{t-1}, q^{t-1})\} \\ &+\{f(U_I^{t-1}, S^t, p^{t-1}, q^{t-1}) - f(U_I^{t-1}, S^{t-1}, p^{t-1}, q^{t-1})\} \end{aligned} \tag{16}$$

According to (15), Lemma 1 and Lemma 2, the three terms at the right side of the above equation are less than or equal to zero. Hence, the sequence $\{F^t\}_{t=1,...}$ is non-increasing. It is easy to verify that F^t is bounded below. Consequently, we can conclude that $\{F^t\}_{t=1,...}$ converges.

For the computational cost of our method, the bottleneck lies on the eigenvalue decomposition of Step 1 in Algorithm 1, where the time complexity is $\mathrm{O}(n^3)$. It can be reduced to $\mathrm{O}(n^{2.376})$ using the Coppersmith-Winograd algorithm. Therefore, the time complexity of the offline training process is $\mathrm{O}(k(d^3 + n^{2.376}))$ approximately, where k is the number of iteration needed to converge, n is the number of training samples, and $d = \max(d1, d2)$, $d1$ and $d2$ are the dimensions of the two modality data, respectively.

4. Experimental Results

Given a cross-modal problem, we can learn two projection matrices on the training set using the iterative algorithm given by Algorithm 1. Then, using the two projection matrices we can project each pair of data into the common subspace defined by class labels, in which the relevance of projected data from different modalities can be easily measured. In the testing phase, we take one modality data of the testing set as the query set to retrieve the other modality data. We apply the proposed LCFS approach to deal with a challenging cross-modal problem, i.e., text-image retrieval. And we evaluate and compare different methods on two publicly available datasets - Pascal VOC 2007 [11] and Wiki image-text dataset [19].

4.1. Experimental settings

We compare the proposed LCFS approach with several related methods, namely, PLS [21], BLM [22, 24], CCA [7, 19], GMMFA and GMLDA [22], for two cross-modal retrieval tasks: (1) Image query vs. Text database, (2) Text query vs. Image database. In testing phase, the cosine distance is adopt to measure the similarity of features. Given an image (or text) query, the goal of each cross-modal task is to find the nearest neighbors from the text (or image) database. We want more correct matches in the top K documents for a better retrieval.

Figure 2. The top nine images retrieved by our method on the Pascal VOC dataset, given the tags "boat+water".

The mean average precision (MAP) [19] is used to evaluate the overall performance of the algorithms. To compute MAP, we first evaluate the average precision (AP) of a set of N retrieved documents by $\mathrm{AP} = \frac{1}{T}\sum_{r=1}^{N} P(r)\delta(r)$, where T is the number of relevant documents in the retrieved set, $P(r)$ denotes the precision of the top r retrieved documents, and $\delta(r) = 1$ if the rth retrieved document is relevant (where relevant means belonging to the class of the query) and $\delta(r) = 0$ otherwise. The MAP is then computed by averaging the AP values over all queries in the query set. Besides, we also use precision-scope curve [18] and precision-recall curve [19] to evaluate the effectiveness of different methods. The scope is specified by the number (K) of top-ranked documents presented to the users. The precision-recall curve is a classical measure of information retrieval performance, but some researchers [18] consider the characterization of retrieval performance by curves of precision-scope more expressive for multimedia retrieval. So we report results with both of the measures here.

For our method, the parameters λ_1 and λ_2 are empirically set to 0.1 and 0.001, respectively, in all experiments.

4.2. Results on Pascal VOC dataset

In this subsection, we first report the results of different methods on the Pascal VOC dataset. The Pascal VOC

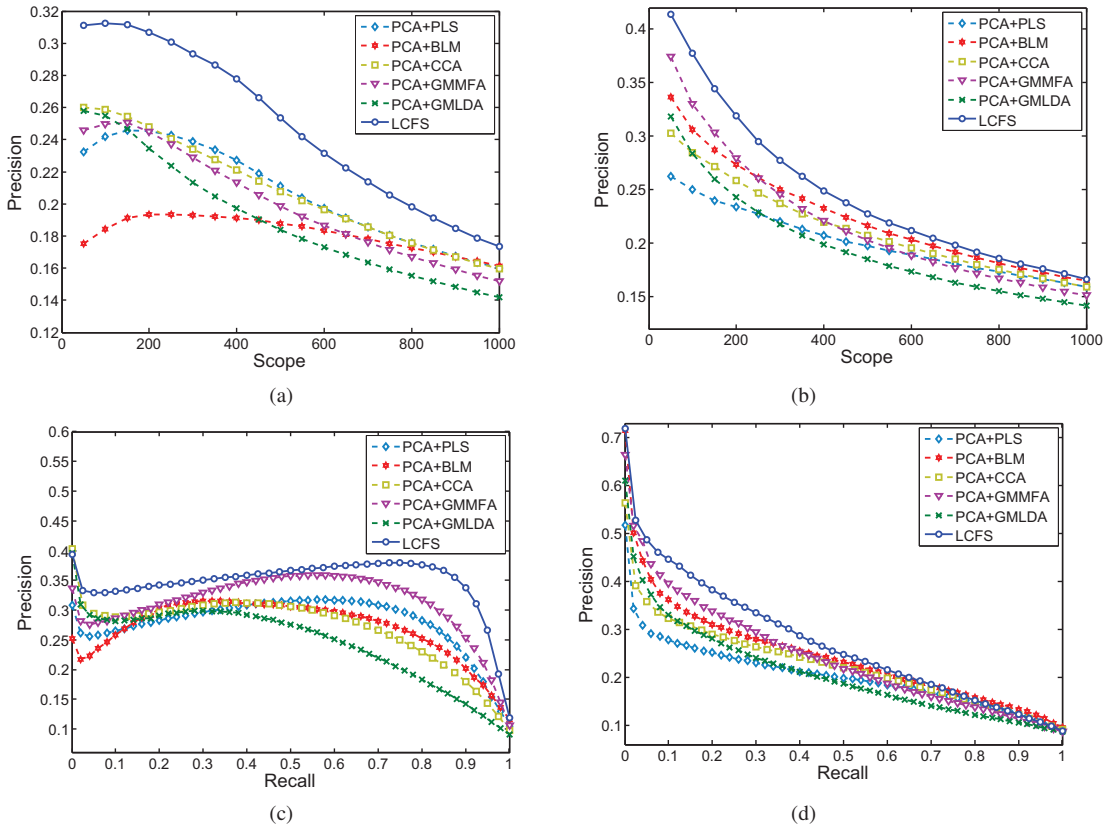

Figure 3. Performance of different methods on the Pascal VOC dataset, based on precision-scope curve (top row) for K = 50 to 1000 and precision-recall curve (bottom row). Left column: Image query vs. Text database. Right column: Text query vs. Image database.

dataset [11] consists of 5011/4952 (training/testing) image-tag pairs, which can be categorized into 20 different classes. Some images are multi-labeled, so we select images with only one object, which results in 2808 training and 2841 testing data. The image features are 512-dimensional Gist features [11], and the text features are 399-dimensional word frequency features.

As we mentioned in Section 2, the compared methods just focus on the common subspace learning, so Principal Component Analysis (PCA) is performed on the original features to remove redundant features. Our method can perform coupled feature selection, so we do not perform PCA on the original features for our method. Table 1 shows the MAP scores achieved by PLS, BLM, CCA, GMMFA, GMLDA and our method (LCFS) on the Pascal VOC dataset. To illustrate the importance of PCA, the results of GMMFA without performing PCA on the original features are also reported, as shown in Table 1, which are much worse than those of performing PCA. We observe that our method outperforms its several counterparts. This may be because our method selects the relevant and discriminative features from the two modalities simultaneously, and the learnt common space is more compact and effective. From Table 1, we also see that GMMFA and GMLDA does not obtain similar results as expected, and GMLDA does not work as well as GMMFA. This may be because the text features of the Pascal VOC dataset are very sparse, which maybe does not agree the assumption of GMLDA.

Figure 2 shows the top nine retrieval images using a tag vector containing "boat+water" as query. Firstly, tag vectors and image feature vectors are projected into the common space by the proposed method. Then, for a tag vector, we return the nearest K images as the retrieved results. We can see that most retrieved images are very relevant to the given query.

The corresponding precision-scope curves and precision-recall curves are plotted in Figure 3. The scope (i.e., the top K retrieved items) for the precision-scope curves varies from K=50 to 1000. The top row shows the performance of different methods based on the precision-scope curves for both forms of cross-modal retrieval tasks, i.e., Image query vs. Text database (left) and Text query vs. Image database (right). We observe that compared with its several counterparts, our method obtains better results for both tasks. The bottom row shows the performance of different methods based on the

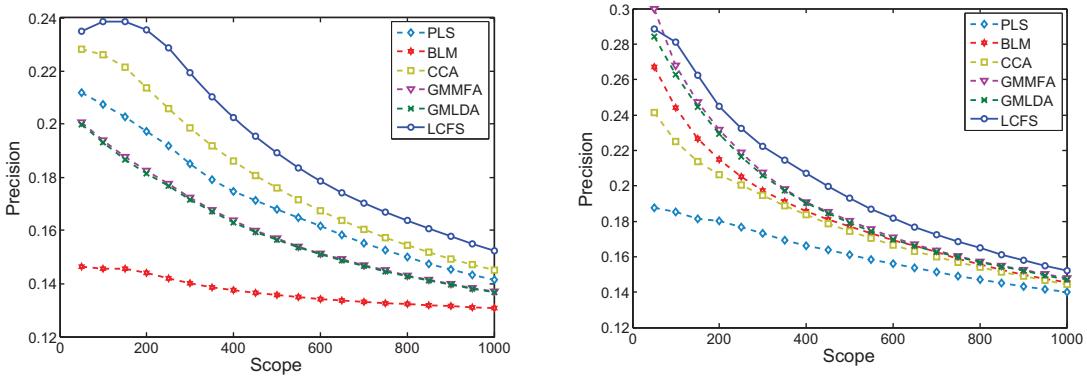

Figure 4. Performance of different methods on the Wiki dataset, based on precision-scope curve for $K = 50$ to 1000. (a) Image query vs. Text database. (b) Text query vs. Image database.

precision-recall curves, and our method also outperforms other methods for both forms of cross-modal retrieval.

4.3. Results on Wiki dataset

The Wiki image-text dataset [19], generated from Wikipedia's "featured articles", consists of 2866 image-text pairs. In each pair, the text is an article describing people, places or some events and the image is closely related to the content of the article. Each pair is labeled with one of 10 semantic classes. We split it into a training set of 1300 pairs (130 pairs per class) and a testing set of 1566 pairs. The representation of the text with 10 dimensions is derived from a latent Dirichlet allocation model. The images are represented by the 128 dimensional SIFT descriptor histograms. Due to the low dimensions of image and text features themselves, PCA is not used to reduce the dimensions of the original features here.

Table 2 shows the MAP scores of different approaches on the Wiki dataset. Our method achieves MAP scores of 0.2798 and 0.2141 for the image query and text query respectively, only a little bit better than those of GMMFA and GMLDA. The reason is that the dimensions of image and text features are low, so the ℓ_{21}-norms of our method for coupled feature selection could hardly take effect. We also see that GMMFA, GMLDA and our method perform better than PLS, BLM, and CCA. This is because BLM, CCA and PLS only use pairwise information, GMMFA, GMLDA and our method use class information, which provides much better separation between classes.

Due to limited space, we only show the corresponding precision-scope curves, which are plotted in Figure 4. We can see that for both forms of cross-modal retrieval, our method finds more correct matches in the top K documents than its compared methods. Figure 5 shows two examples of text queries and the top five images retrieved by our method. In each case, the query text and its paired image are shown at the left, and the top five images are shown at columns 3-7. Note that our method finds the closest matches at semantic level, i.e., the common space defined by class labels. The retrieved images are perceived as belonging to the same category of the query text ("Geography & places" at the top, "Warfare" at the bottom).

Methods	Image query	Text query	Average
PLS	0.2402	0.1633	0.2032
BLM	0.2562	0.2023	0.2293
CCA	0.2549	0.1846	0.2198
GMMFA	0.2750	0.2139	0.2445
GMLDA	0.2751	0.2098	0.2425
LCFS	**0.2798**	**0.2141**	**0.2470**

Table 2. Comparison of MAP (Mean Average Precision) performance of different methods on the Wiki dataset.

5. Conclusion

In this paper, we have proposed a general regularization framework to solve the problem of cross-modal matching, which consists of coupled subspace learning for different modalities, the ℓ_{21}-norms for coupled feature selection , and the trace norm for the measurement of relevance. Under the framework, different projection matrices are learnt to project different modal data into a common subspace defined by label information, and relevant and discriminative features for the coupled spaces are selected simultaneously in the projection procedure. To solve this complex regularization problem, we have harnessed an alternative formulation of the trace norm, and reformulated ℓ_{21}-norm based on half-quadratic analysis, which leads to an iterative algorithm. Experimental results on two public cross-modal datasets have demonstrated that the proposed method performs better than some relevant state-of-the-art methods.

Figure 5. Two examples of text queries (the first column) and the top five images (columns 3-7) retrieved by our method on the Wiki dataset. The second column contains the paired images of the text queries .

Acknowledgment

This work is jointly supported by National Basic Research Program of China (2012CB316300), National Natural Science Foundation of China (61175003, 61135002, 61202328, 61103155), and Hundred Talents Program of CAS.

References

[1] R. Angst, C. Zach, and M. Pollefeys. The generalized trace-norm and its application to structure-from-motion problems. *In ICCV*, pages 2502–2509, 2011.

[2] Y. Chen, L. Wang, W. Wang, and Z. Zhang. Continuum regression for cross-modal multimedia retrieval. *In ICIP*, pages 1949–1952, 2012.

[3] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimizaition. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

[4] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. *In NIPS*, pages 2187–2195, 2011.

[5] Q. Gu, Z. Li, and J. Han. Joint feature selecion and subspace learning. *In IJCAI*, pages 1294–1299, 2011.

[6] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale image classification with trace-norm regularization. *In CVPR*, pages 3386–3393, 2012.

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

[8] R. He, T. N. Tan, L. Wang, and W. Zheng. ℓ_{21} regularized correntropy for robust feature selection. *In CVPR*, pages 2504–2511, 2012.

[9] R. He, W. Zheng, and B. Hu. Maximum correntropy criterion for robust face recognition. *IEEE TPAMI*, 33(8):1561–1576, 2011.

[10] Z. Huang, S. Shan, H. Zhang, S. Lao, and X. Chen. Cross-view graph embedding. *In ACCV*, 2012.

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

[12] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. *In CVPR*, pages 1123–1128, 2009.

[13] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. *In CVPR*, pages 605–611, 2009.

[14] A. Li, S. Shan, X. Chen, and W. Gao. Face recognition based on non-corresponding region matching. *In ICCV*, pages 1060–1067, 2011.

[15] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selction via joint ℓ_{21}-norms minimization. *In NIPS*, pages 1813–1821, 2010.

[16] M. Nikolova and M.K.Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific Computing*, 27(3):937–966, 2005.

[17] N. Quadrianto and C. H. Lampert. Learning multi-view neighborhood preserving projections. *In ICML*, pages 425–432, 2011.

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

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

[20] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *LNCS*, pages 34–51, 2006.

[21] A. Sharma and D. W. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. *In CVPR*, pages 593–600, 2011.

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

[23] L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. *In ICML*, pages 1024–1031, 2008.

[24] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[25] R. Udupa and M. Khapra. Improving the multilingual user experience of wikipedia using cross-language name search. *NACACL-HLT*, pages 492–500, 2010.

[26] F. Wu, Y. Yuan, X. Liu, J. Shao, Y. Zhuang, and Z. Zhang. The heterogeneous feature selection with structual sparsity for multimedia annotation and hashing: a survey. *IJMIR*, 1(1):3–15, 2012.