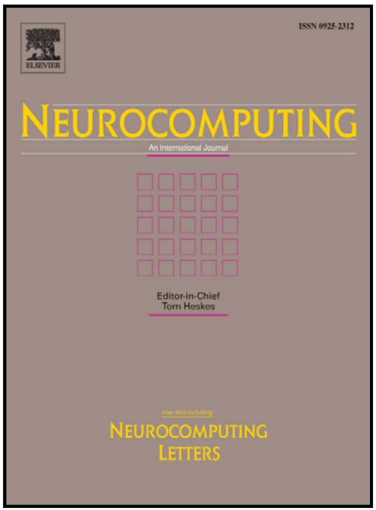


www.elsevier.com/locate/neucom

Multi-view Multi-sparsity Kernel Reconstruction for Multi-class Image Classification

Xiaofeng Zhu, Qing Xie, Yonghua Zhu, Xingyi Liu, Shichao Zhang

PII: S0925-2312(15)00685-2
DOI: http://dx.doi.org/10.1016/j.neucom.2014.08.106
Reference: NEUCOM15550

To appear in: *Neurocomputing*

Received date: 29 April 2014
Revised date: 19 August 2014
Accepted date: 25 August 2014

Multi-view Multi-sparsity Kernel Reconstruction for Multi-class Image Classification

Xiaofeng Zhu[1,2], Qing Xie[3], Yonghua Zhu[4], Xingyi Liu[5], Shichao Zhang[2*]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, P. R. China*
[2] *Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, P. R. China*
[3] *Division of CEMSE, KAUST, Saudi Arabia*
[4] *School of Computer, Electronics and Information, Guangxi University, China*
[5] *Qinzhou Institute of Socialism, Qinzhou, Guangxi, China*

Abstract

This paper addresses the problem of multi-class image classification by proposing a novel multi-view multi-sparsity kernel reconstruction (MMKR for short) model. Given images (including test images and training images) representing with multiple visual features, the MMKR first maps them into a high-dimensional space, e.g., a reproducing kernel Hilbert space (RKHS), where test images are then linearly reconstructed by some representative training images, rather than all of them. Furthermore a classification rule is proposed to classify test images. Experimental results on real datasets show the effectiveness of the proposed MMKR while comparing to state-of-the-art algorithms.

Keywords: image classification, multi-view classification, sparse coding, Structure sparsity, Reproducing kernel Hilbert space

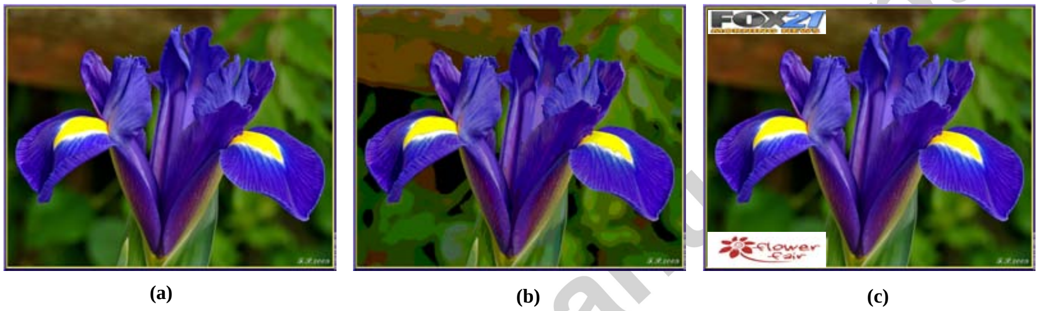
1. Introduction

In image classification, an image is often represented by its visual feature, such as HSV (Hue, Saturation, Value) color histogram, LBP (Local Binary Pattern), SIFT (Scale invariant feature transform), CENTRIST (CENsus TRansform hISTgram), and so on. Usually, different representations describe different char-

[*] Corresponding author.
Email address: zhangsc@mailbox.gxnu.edu.cn (Shichao Zhang[2])

acteristics of images. For example, CENTRIST [32] is a suitable representation for place and scene recognition.

Recent studies (e.g., [32]) have shown that although an optimal representation (such as SIFT) is better for some given tasks, it might no longer be optimal for the others. Moreover, a single visual feature is not always robust to all types of scenarios. Give an example illustrated in Figure 2, we can easily classify the two figures (e.g., Figure 2.(a) and 2.(b)) into the category IRIS according to the extracted local feature. However, we maybe not easily make the same decision while giving their global feature, such as HSV. Actually, in this case, we may category two figures (e.g., Figure 2.(a) and 2.(c)) into IRIS. According to our observation, we cannot category Figure 2.(c) into IRIS since the captions in Figure 2.(c) makes the classification difficult.



(a) (b) (c)

Figure 1: A illustration on image IRIS with different representations.

In contrast, literatures (e.g., [41, 39]) have shown that representing image data with multiple features really reflects the specific information of image data. Moreover, this case is complementary each other and helpful for disambiguation. For example, the local feature HSV is less robust to the changes in frame rate, video length, captions. SIFT is sensitive to changes in contrast, brightness, scale, rotation, camera viewpoint, and so on [13, 42].

Aforementioned observation motivates us to combine several visual features (rather than a single type of visual feature) to perform image visual classification for discriminating each class best from all other classes. In the machine learning domain, learning with multiple representations is well known as multi-view learning (MVL) or multi-modality learning [2].

Using multi-view learning brings clear advantages over traditional single-view learning: First, multi-view learning is more effective than generating a single model via considering all attributes at once, especially when the weaknesses of one view complement the strengths of the others [7]. In many application areas, such as bioinformatics and video summarization, literatures have shown that multimedia classification can achieve greatly benefit from multi-view learning [21, 23]. Second, different information about the same example in multi-view learning can help solve other issues, such as transfer learning and semi-supervised learning [32]. Therefore, multi-view learning is becoming popular in real applications [11, 16, 33], such as web analysis, object recognition, image classification, and so on.

However, previous studies on multi-view learning contain at least two following drawbacks. First, multi-view learning employ all the views for each data point without considering the individual characteristics of each data point. For example, sometimes a data point can be described well with several representations and can not be added any other. In this case, we really expect to select the best suitable views according to its characteristics. Second, in real application, image datasets are often corrupted by noise, but existing multi-view learning approaches have difficulty for dealing with noisy observations [6]. Therefore, we expect to remove the noise or redundancy from the training data for selecting appropriated views for each image.

In this paper we extend our previous work [43][1] to conduct multi-class image classification by proposing a multi-view multi-sparsity kernel reconstruction (MMKR) model. Specifically, the MMKR performs kernel reconstruction in a RKHS, in which each test image is linearly reconstructed by training images coming from a few object categories, via a new designed multi-sparsity regularizer, which concatenates an ℓ_1-norm with a Frobenius norm (F-norm for short) for achieving following advantages, such as selecting training images from a few object categories to reconstruct the test image via the F-norm regularizer, and removing noise in visual features via the ℓ_1-norm regularizer. Finally, experimental results on challenging real datasets show the effectiveness of the proposed MMKR to the state-of-the-art algorithms.

The remainder of the paper is organized as below: Preliminary is described in Section 2, followed by the proposed MMKR approach in Section 3 and its optimization in Section 4. The experimental results are reported and analyzed in Section 5 while Section 6 concludes the paper.

[1] Different from our conference version [43], this paper added the Related Work, rewrote the Introduction, and revised the parts, such as Approach and Experimental Analysis.

2. Related work

In this section, we give a brief review on multi-view learning and spare learning.

2.1. Multi-view learning

Multi-view learning learns one task of the data with multiple visual features. The basic idea of multi-view learning is to make use of the consistency among different views to achieve better performance. Many literatures (e.g., [12, 26]) showed that multi-view learning can improve learning performance in all kinds of real applications, such as natural language tasks, computer vision, and so on [8, 16].

The study in [2] may be the earliest work on multi-view learning, where the authors proposed a co-training approach to learn the data described by two distinct views. Recently, Chaudhuri et al. [5] employed canonical correlation analysis (CCA) to perform clustering and regression in multi-view learning. Chen et al. [6] proposed a large-margin framework for learning multi-view data.

In multi-view learning, the information in some a view can help to solve the weakness of the other views, so multi-view learning has been embedded into many types of learning tasks, such as semi-supervised multi-view learning and transfer multi-view learning. For example, the literatures (e.g., [35]) found that each view should follow same data distribution in semi-supervised learning, but their proposed semi-supervised multi-view learning can be used to more flexible cases, i.e., views can follow different data distribution each other. Moreover, they incorporated the consistency among views to perform semi-supervised multi-view learning. Finally, they showed that their proposed semi-supervised multi-view learning is with a substantial improvement on the classification performance than existing methods. In transfer multi-view learning, the literatures (e.g., [4, 38]) leveraged the consistency of the views and considered the domain difference among the views to learn heterogenous data.

2.2. Sparse learning

The objective function of traditional sparse learning can be represented as the following form:

$$\min_{\text{parameters}} \text{loss function} + \text{regularizer} \tag{1}$$

Loss function in Eq.1 is used to achieve minimal regression (or reconstruction) error. Existing loss functions include least square loss function, logistic loss function, squared hinge loss function, and so on. The regularizer is often used to meet some goals, such as avoiding the issue of over-fitting, leading to the sparsity, and so on. In real applications, sparse learning has been applied in reconstruction process (e.g., [9, 20]) or regression process (e.g., [14, 27]).

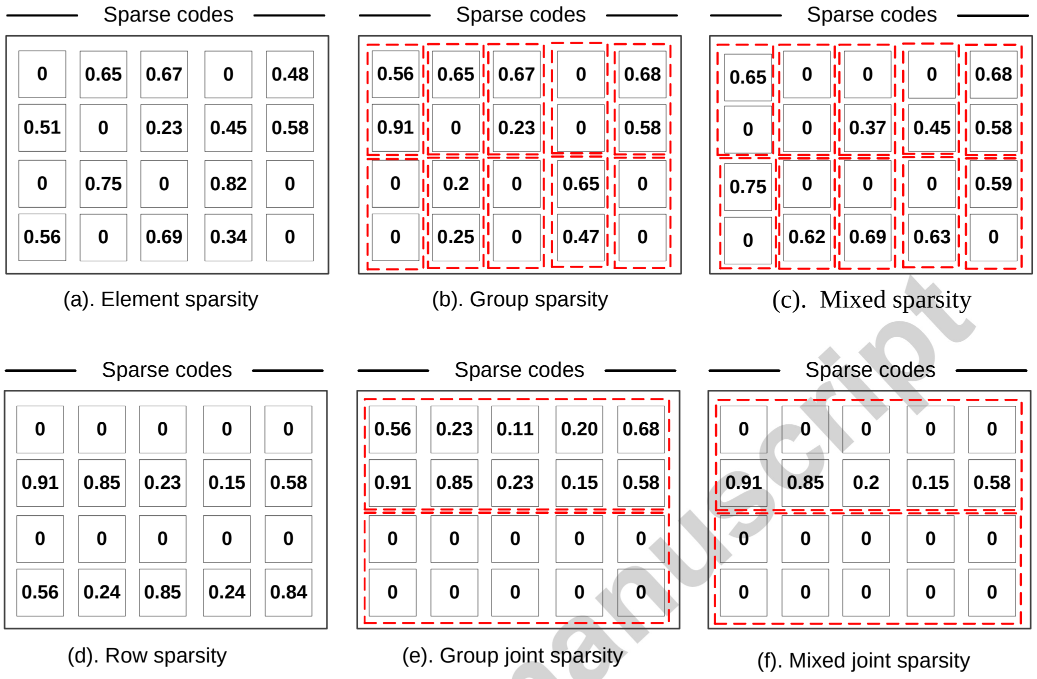


Figure 2: An illustration on different types of sparsity in separable sparse learning (i.e., the left three subfigures) and joint sparse learning (i.e., the right three subfigures). Note that, a red box means one group. For better viewing, please see the original color pdf file.

Sparse learning codes a sample (e.g., a signal) using a few number of dictionaries (or atoms in signal analysis) via the form in Eq.1. The key idea of sparse learning is to generate sparse results, which makes the learning more efficient [40]. The literatures (e.g., [14, 27]) showed different regularizers encourage various sparsity pattern in sparse learning. According to the way to generate sparsity patterns, we categorize existing sparse learning into two parts, i.e., separable sparse learning (e.g., [9, 36], or please see examples form Fig.2.(a) to Fig.2.(c)) and joint sparse learning (e.g.,[1, 31, 34], or please see examples form Fig.2.(d) to Fig.2.(f)) respectively. Separable sparse learning codes one sample once. Joint sparse learning model simultaneously codes all samples. For example, there are four samples in each subfigure of Fig.2, each column is the sparse codes of one sample. To generate sparse codes for all four samples, separable sparse learning needs to perform its optimization process four times. However, joint sparse learning only needs one time.

Separable sparse learning employs different regularizers to lead to different sparse patterns. For example, the ℓ_1-norm regularizer (e.g., [9]) leads to the element sparsity; the $\ell_{2,1}$-norm regularizer (e.g., [36]) for the group sparsity and the mixed-norm regularizer (concatenating a ℓ_1-norm regularizer with a $\ell_{2,1}$-norm regularizer, e.g.,[22]) for the mixed sparsity. To generate the sparsity, the ℓ_1-norm regularizer makes each code as a singleton, then generates four codes in the first column of Fig.2.(a) independently. The ℓ_1-norm regularizer also generates codes of each sample in Fig.2.(a) independently. The resulted sparsity is called as element sparsity. The groups sparsity is obtained by forcing a group in one column as a singleton, so its sparsity is generated in the whole group, e.g., the second red box (i.e., group) in the first column of Fig.2.(b). Obviously, the $\ell_{2,1}$-norm regularizer inducing the group sparsity takes the natural group structure in one example into account. However, it still generates sparse codes one sample once. The mixed sparsity has been explained (e.g., [22]) as first generating the group sparsity for each sample, e.g., sparsity in the second red box (i.e., group) in the first column of Fig.2.(c), and then generating the element sparsity in the dense (i.e., non-sparse) groups, e.g., the second element in the first column of Fig.2.(c). In a word, although the mixed sparsity hierarchically generates the group sparsity and the element sparsity, it is generated one sample once.

The regularizers (e.g.,the $\ell_{2,1}$-norm regularizer (e.g., [28, 31]), the $\ell^2_{2,1}$-norm regularizer (e.g., [1]), the $\ell_{1,\infty}$-norm regularizer (e.g., [24])) are often used in joint sparse learning. Different from generating sparse codes one sample once in separable sparse learning, joint sparse learning considers to simultaneously encode all samples (i.e., all four sample in Fig.2) by requiring them to share same dictionaries. For example, the row sparsity (via the $\ell_{2,1}$-norm regularizer) in Fig.2.(d) enables all four samples to be encoded at the same time, and the sparsity through the whole row, such as the first row and the third row. The block sparsity via the F-norm regularizer considers the natural group structure, i.e., first two rows as one block, and the last two rows as another block, so it generate the sparsity through the whole block, e.g., the second block in Fig.2.(e).

3. Approach

In this paper, we denote matrices as boldface uppercase letters, vectors as boldface lowercase letters, and scalars as normal italic letters, respectively. For a matrix $\mathbf{X} = [x_{ij}]$, its *i*-th row and *j*-th column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively. Also, we denote the Frobenius norm and $\ell_{2,1}$-norm of a matrix $\mathbf{X}$ as $\|\mathbf{X}\|_F = \sqrt{\sum_i \|\mathbf{x}^i\|_2^2} = \sqrt{\sum_j \|\mathbf{x}_j\|_2^2}$ and $\|\mathbf{X}\|_{2,1} = \sum_i \|\mathbf{x}^i\|_2 = \sum_i \sqrt{\sum_j x_{ij}^2}$, respectively. We further denote the transpose operator, the trace operator, and the inverse of a matrix $\mathbf{X}$ as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

Given a set of training images $\mathbf{X}$, each image is represented with V visual features (or views) and described by one of C object categories appeared in $\mathbf{X}$. We denote $\mathbf{x}_c^v$ ($\mathbf{x}_c^v \in R^{m_v}, c = 1, ..., C, v = 1, ..., V$) as the v-th view of an image in c-th object category. $\mathbf{X}_c^v$ ($\mathbf{X}_c^v \in R^{m_v \times n_c}$) is a set of training images associated with the c-th object category and represented by the v-th view. We also denote $\sum\limits_{v=1}^{V} m_v = M$, and $\sum\limits_{c=1}^{C} n_c = N$, where N is training size and m_v is the dimensionality of the v-th view.

3.1. Objective function

Sparse learning distinguishes important elements from unimportant ones by assigning the codes of unimportant elements as zero and the important ones as non-zero. This enables that sparse learning reduces the impact of noises and increase the efficiency of learning models [17]. Thus it has been embedded into various learning models, such as sparse principal component analysis (sparse PCA [44]), sparse non-negative matrix factorization (sparse NMF [15]), and sparse support vector machine (sparse SVM [29]), in many real applications [3, 13], including signal classification, face recognition and image analysis [30]. In this paper, we cast multi-view image classification as multi-view sparse learning in the RKHS.

Given the v-th visual feature of a test image $\mathbf{y}^v$ ($\mathbf{y}^v \in R^{m_v}$), we first search for a linear relationship between $\mathbf{y}^v$ and the v-th visual feature of training images. For this, we consider to build a reconstruction process as: $f(\mathbf{y}^v) = \sum\limits_{c=1}^{C} \mathbf{y}_c^v \mathbf{w}_c^v$, where $\mathbf{w}_c^v \in R^{n_c}$ is the v-th view reconstruction coefficient.

To perform the reconstruction process in multi-view learning, we expect to minimize reconstruction error across all the views. To avoid the issue of overfitting as well as to obtain sparse effect, we propose a regularizer leading to multiple sparsity into the framework of sparse learning, i.e., the proposed multi-sparsity regularizer includes an ℓ_1-norm and an F-norm for achieving the element sparsity and the block sparsity respectively. The objective function is defined as:

$$\min_{\mathbf{w}_c^1,\dots,\mathbf{w}_c^V} \quad \sum_{v=1}^{V} \|\mathbf{y}^v - \sum_{c=1}^{C} \mathbf{X}_c^v \mathbf{w}_c^v\|_2^2 + \lambda_1 \sum_{v=1}^{V} \sum_{c=1}^{C} |w_c^v| + \lambda_2 \sum_{c=1}^{C} \sqrt{\sum_{v=1}^{V} (w_c^v)^2} \qquad (2)$$

where λ_1 and λ_2 are trade-off parameters. The first term in Eq.2 is to minimize the reconstruction error through all views. The last two terms are introduced to avoid the issue of over-fitting and to pursue multi-sparsity.

For convenience, we denote: $\tilde{\mathbf{x}}_c^v = [0, ..., 0, (\mathbf{x}_c^v)^T, 0, ..., 0]^T$, $\tilde{\mathbf{y}}^v = [0, ..., 0, (\mathbf{y}^v)^T, \ 0, ..., 0]^T$, $\tilde{\mathbf{w}}_c^v = [0, ..., 0, (\mathbf{w}_c^v)^T, 0, ..., 0]^T$, and $\tilde{\mathbf{W}} = [(\tilde{\mathbf{w}}_1)^T, ..., (\tilde{\mathbf{w}}_C)^T]^T \in R^{M \times V}$, where $\tilde{\mathbf{w}}_c \in R^{n_c \times V}$, where both $\tilde{\mathbf{x}}_c^v (\in R^M)$ and $\tilde{\mathbf{y}}^v (\in R^M)$ are a one-dimensional column vector with the $(\sum_{i=1}^{v-1} m_i + 1)$-th to the $\sum_{i=1}^{v} m_i$-th elements being nonzero. Therefore, Eq.2 can be converted as:

$$\min_{\tilde{\mathbf{W}}} \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\mathbf{W}}\|_F^2 + \lambda_1 \|\tilde{\mathbf{W}}\|_1 + \lambda_2 \sum_{c=1}^{C} \|\tilde{\mathbf{W}}_c\|_F \qquad (3)$$

where $\|.\|_F$ denotes F norm, $\tilde{\mathbf{X}}_c \in R^{M \times n_c}$, $\tilde{\mathbf{Y}} \in R^{M \times V}$ and $\tilde{\mathbf{Y}} \in R^{n_c \times V}$.

However, Eq.3 is developed for image classification in original space. Motivated by the fact that kernel trick can capture nonlinear similarity, which has been demonstrated to reduce feature quantization error and boost learning performance, we use a nonlinear function ϕ^v in each view v to map training images and test images from original space to a high-dimensional space, e.g., the RKHS, via defining $k(x_i, x_j)^v = \phi(x_i^v)^T \phi(x_j^v)$ for some given kernel functions k^v, where $v = 1, ..., V$. That is, given a feature mapping function $\phi : R^M \to R^K$, ($M < K$), both training images and test images in feature space R^M are mapped into a RKHS R^K via ϕ, i.e., $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_M] \to \phi(\tilde{\mathbf{X}}) = [\phi(\tilde{\mathbf{x}}_1, ..., \phi(\tilde{\mathbf{x}}_M)]$. By denoting $\mathbf{A} = \phi(\tilde{\mathbf{X}})^T \phi(\tilde{\mathbf{Y}})$ and $\mathbf{B} = \phi(\tilde{\mathbf{X}})^T \phi(\tilde{\mathbf{X}})$, we convert the objective function defined in the original space (see eq.3) to the objective function of the proposed MMKR as:

$$\min_{\tilde{\mathbf{W}}} \|\mathbf{A} - \mathbf{B}\tilde{\mathbf{W}}\|_F^2 + \lambda_1 \|\tilde{\mathbf{W}}\|_1 + \lambda_2 \sum_{c=1}^{C} \|\tilde{\mathbf{W}}_c\|_F \qquad (4)$$

where $A \in R^{K \times V}$ and $B \in R^{K \times N}$.

According to the literatures, e.g.,[14], the λ_1-norm regularizer generates the element sparsity, whose sparsity is in single element of $\tilde{\mathbf{W}}$, and benefits for removing noise by assigning its codes as sparse, i.e., 0. The F-norm regularizer generates the block sparsity, whose sparsity is through the whole block, i.e., zero through the whole object category in this paper. Thus the F-norm regularizer enables the object categories with the block sparsity (i.e., sparsity in each code through the whole objective category) not to be involved into the reconstruction process. By inducing the multi-sparsity regularizer, only a few training images from representative object categories are used to reconstruct each test image. Meanwhile, removing noise is also considered.

3.2. Classification rule

By solving the objective function in Eq.4, we obtain the optimal $\tilde{\mathbf{W}}$. According to the literature in [37], for each view v, if we use only the optimal coefficients $\mathbf{W}_c^v$ associated with the c-th class, we can approximate the v-th view $\mathbf{y}^v$ of the test image as $\phi(\mathbf{y}^v) = \phi(\mathbf{X}_c^v)W_c^v$. Then the classification rule is defined as in favor of the class with the lowest total reconstruction error through all the V views: where θ_v, ($c = 1, ..., V$ and $\sum\limits_{v=1}^{V} \theta_v = 1$) is the weight measuring the confidence of the v-th view in the final decision. We only simply set $\theta_v = \frac{1}{V}$ in this paper.

4. Optimization

Eq.4 is convex, so it admits the global optimum. However, its optimization is very challengeable because both the $\|\tilde{\mathbf{W}}\|_F$-norm and the $\|\tilde{\mathbf{W}}\|_1$-norm in Eq.4 are convex but non-smooth. In this section we propose a simple algorithm to optimize Eq.4.

By setting the derivative of Eq.4 with respect to $\tilde{\mathbf{w}}_i$ ($1 \leq i \leq V$) as zero, we obtain:

$$(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{E}_i + \lambda_2\mathbf{D})\tilde{\mathbf{w}}_i = \mathbf{B}^T\mathbf{a}_i \tag{5}$$

where $\mathbf{E}_i$ is a diagonal matrix with the k-th diagonal element as $\frac{1}{2|\tilde{w}_k^i|}$ and $\mathbf{A} = \{\mathbf{a}_1, ..., \mathbf{a}_V\}$. $\mathbf{D} = diag(\mathbf{D}_1, ..., \mathbf{D}_C)$, the '*diag*' is the diagonal operator and each $\mathbf{D}_c$ ($c = 1, ..., C$) is also a diagonal matrix with the j-th diagonal element as $D_{j,j} = \frac{1}{2\|\tilde{\mathbf{w}}_c\|_F}$, $j = 1, ..., n_c$.

By observing Eq.5, we find that both $\mathbf{E}_i$ and $\mathbf{D}$ depend on the value of $\tilde{\mathbf{W}}$. In this paper, following the literatures [18, 42], we design a novel iterative algorithm (i.e., Algorithm 1) to optimize Eq.4 and then prove its convergence. Here we introduce Theorem 1 to guarantee that Eq.4 monotonically decreases in each iteration of Algorithm 1.

We first give a lemma as follows:

Lemma 1. *For any positive values α_i and β_i, $i = 1, ..., m$, the following holds:*

$$\begin{aligned}&\sum_{i=1}^{m} \frac{\beta_i^2}{\alpha_i} \leq \sum_{i=1}^{m} \frac{\alpha_i^2}{\alpha_i} \Longleftrightarrow \sum_{i=1}^{m} \frac{(\beta_i+\alpha_i)(\beta_i-\alpha_i)}{\alpha_i} \leq 0 \\ &\Longleftrightarrow \sum_{i=1}^{m}(\beta_i - \alpha_i) \leq 0 \Longleftrightarrow \sum_{i=1}^{m} \beta_i \leq \sum_{i=1}^{m} \alpha_i\end{aligned} \tag{6}$$

Theorem 1. *In each iteration, Algorithm 1 monotonically decreases the objective function value in Eq.4.*

Proof. According to the sixth step of Algorithm 1, we denote $\mathbf{W}^{[t+1]}$ as the results of the $(t+1)$-th iteration of Algorithm 1, then we have:

$$\begin{aligned}\tilde{\mathbf{W}}^{[t+1]} = \min_{\tilde{\mathbf{W}}} \quad & \frac{1}{2}\|\mathbf{A} - \mathbf{B}\tilde{\mathbf{W}}\|_F^2 + \lambda_1 \sum_{i=1}^{V} \tilde{\mathbf{W}}_i^T E_i^{[t]} \tilde{\mathbf{W}}_i \\ & + \lambda_2 \sum_{c=1}^{C} tr((\tilde{\mathbf{W}}_c)^T (\mathbf{D}_c)^{[t]} \tilde{\mathbf{W}}_c)\end{aligned} \tag{7}$$

then we can obtain:

$$\begin{aligned}&\frac{1}{2}\|\mathbf{A} - \mathbf{B}(\tilde{\mathbf{W}}^{[t+1]})^T\|_F^2 + \lambda_1 \sum_{i=1}^{V} \tilde{\mathbf{W}}_i^T E_i^{[t]} \tilde{\mathbf{W}}_i \\ &\quad + \lambda_2 \sum_{c=1}^{C} tr((\tilde{\mathbf{W}}_c)^T (\mathbf{D}_c)^{[t]} \tilde{\mathbf{W}}_c) \\ &\leq \frac{1}{2}\|\mathbf{A} - \mathbf{B}(\tilde{\mathbf{W}}^{[t]})^T\|_F^2 + \lambda_1 \sum_{i=1}^{V} \tilde{\mathbf{W}}_i^T E_i^{[t]} \tilde{\mathbf{W}}_i \\ &\quad + \lambda_2 \sum_{c=1}^{C} tr((\tilde{\mathbf{W}}_c)^T (\mathbf{D}_c)^{[t]} \tilde{\mathbf{W}}_c)\end{aligned} \tag{8}$$

which indicates that:

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{A}-\mathbf{B}(\tilde{\mathbf{W}}^{[t+1]})^T\|_F^2 + \lambda_1 \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{((\tilde{w}_i^j)^{[t+1]})^2}{2\|(\tilde{w}_i^j)^{[t]}\|_2} \\
&\qquad + \lambda_2 \sum_{c=1}^{C} \frac{\|(\tilde{\mathbf{W}}_c)^{[t+1]}\|_F^2}{2\|(\tilde{\mathbf{W}}_c)^{[t]}\|_F} \\
&\le \frac{1}{2}\|\mathbf{A}-\mathbf{B}(\tilde{\mathbf{W}}^{[t]})^T\|_F^2 + \lambda_1 \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{((\tilde{w}_i^j)^{[t]})^2}{2\|(\tilde{w}_i^j)^{[t]}\|_2} \\
&\qquad + \lambda_2 \sum_{c=1}^{C} \frac{\|(\tilde{\mathbf{W}}_c)^{[t]}\|_F^2}{2\|(\tilde{\mathbf{W}}_c)^{[t]}\|_F}
\end{aligned} \tag{9}$$

Substituting β_i and α_i with $((\tilde{w}_i^j)^{[t+1]})^2$ (or $\|(\tilde{\mathbf{W}}_c)^{[t+1]}\|_F$) and $((\tilde{w}_i^j)^{[t]})^2$ (or $\|(\tilde{\mathbf{W}}_c)^{[t]}\|_F$) in Lemma 1, we have:

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{A}-\mathbf{B}(\tilde{\mathbf{W}}^{[t+1]})^T\|_F^2 + \lambda_1\|\tilde{\mathbf{W}}^{[t+1]}\|_1 + \lambda_2 \sum_{c=1}^{C} \|(\tilde{\mathbf{W}}_c)^{[t+1]}\|_F \\
&\le \frac{1}{2}\|\mathbf{A}-\mathbf{B}(\tilde{\mathbf{W}}^{[t]})^T\|_F^2 + \lambda_1\|\tilde{\mathbf{W}}^{[t]}\|_1 + \lambda_2 \sum_{c=1}^{C} \|(\tilde{\mathbf{W}}_c)^{[t]}\|_F^2
\end{aligned} \tag{10}$$

This indicates that Eq.4 monotonically decreases in each iteration of Algorithm 1. Therefore, due to the convexity of Eq.4, Algorithm 1 can enable Eq.4 to converge to its global optimum. □

To evaluate the effectiveness of the proposed MMKR, we apply it and several state-of-the-art methods to multi-class object categorization on real datasets [19], such as 17 category and Caltech101 respectively. The comparison algorithms include KMTJSRC [37] only considering the block sparsity in RKHS, KSR [10] only considering the element sparsity in RKHS, the representatives of multiple kernel learning (MKL) methods, e.g., [25].

In our experiments, we obtain kernel matrices by computing $exp(-\chi^2(x,x')/\mu$, where μ is set to be the mean value of the pairwise χ^2 distance on training set.

In the following parts, first, we test parameters' sensitivity of the proposed MMKR according to the variation on parameters λ_1 and λ_2 in Eq.4, aiming at

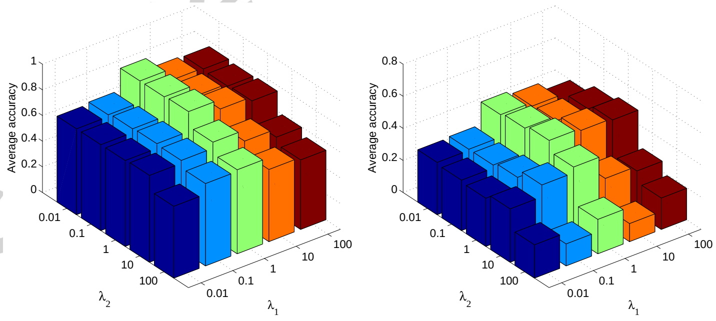
Algorithm 1: The proposed method for solving Eq.4.

```
Input: $\mathbf{A}, \mathbf{B}, \lambda_1$ and $\lambda_2$;
Output: $\tilde{\mathbf{W}} \in R^{N \times V}$;
Initialize $t$ = 1;$\tilde{\mathbf{W}}^{[1]}$ ;
repeat
    Update the $k$-th element in the diagonal matric $\mathbf{E}_i^{[t+1]}$ via $\frac{1}{2|(\tilde{w}_k^i)^{[t]}|}$;
    Update the $c$-th diagonal matrix in the diagonal matrix $\mathbf{D}^{[t+1]}$ via $(D_{j,j})^{[t]} = \frac{1}{2\|(\tilde{\mathbf{W}}_c)^{[t]}\|_F}$;
    for each i,1 $\leq i \leq C$,
        $\tilde{\mathbf{W}}_i^{[t+1]} = \left(\mathbf{B}^T\mathbf{B} + \lambda_1\mathbf{E}_i^{[t]} + \lambda_2\mathbf{D}^{[t]}\right)^{-1}\mathbf{B}^T\mathbf{a}_i$;
    $t$ = $t$+1;
until No change on the objective function value in Eq.4;
```

achieving its best performance. Second, we compare the MMKR with comparison algorithms in terms of average accuracy, i.e., classification accuracy averaged over all classes.

4.1. Parameters' sensitivity

In this subsection we test different settings on parameters (i.e., λ_1 and λ_2 in Eq.4) in our proposed model, and set the value of them varying as $\{0.01, 0.1, 1, 10, 100\}$. The performance on average accuracy of the MMKR is illustrated in Fig.2.



(a) Flower

(b) Caltech

Figure 3: Average accuracy on various parameters' setting at different datasets.

Table 1: Average accuracy (mean±standard deviation)on all algorithms at different datasets. Note that the best results are emphasized through bold-face.

Method	**Flower**	**Caltech**
KSR	0.6301±0.0308	0.4063±0.07545
MKL	0.7460±0.0171	0.4674±0.03956
KMTJSRC	0.7522±0.0336	0.4856±0.05952
MMKR	**0.8022±0.0357**	**0.5124±0.03457**

From Figure 3, we also find the best performance is always obtained in cases with moderate value on both the λ_1 and the λ_2. For example, while the value of parameters' pair (λ_1, λ_2) is (1, 1) for both dataset Flower and dataset Caltech, our MMKR achieves the best average accuracy. Actually, according to our experiments, these cases lead to both the element sparsity (via the λ_1) and the block sparsity (via the λ_2). This illustrates it is feasible to select some training images from a few object categories to perform multi-class image classification.

4.2. Results

In this subsection, we set the values of parameters for the compared algorithms by following the instructions in [37]. For all the algorithms, we repeated each sample ten runs. We recorded the best performance on each combination of their parameters' setting in each run, and reported average results and the corresponding standard deviation in ten runs. The results were illustrated in Table 1.

From Table 1, we can make our conclusions as: 1) The proposed MMKR achieved the best performance. It illustrated that our MMKR was the most effective for multi-class image classification in our experiments. This occurred because the MMKR performed multi-class image classification via deleting noise in training data as well as representing the test image with only some training images from a few object categories. 2) The KMTJSRC outperformed traditional multiple kernel learning methods. This conclusion was consistent to the ones in the literature [37]. 3) Both the proposed MMKR and the KMIJSRC outperformed the KSR because the former two methods reconstructed the test image with some training images, rather than using all training images used in the KSR.

5. Conclusion

In this paper we have addressed the issue of multi-class image classification by first mapping the images (including training images and test images) into a

RKHS. In the RKHS, each test image was linearly reconstructed with training images from a few object categories. Meanwhile, removing noise was also considered. Then a classification rule was proposed by considering the derived reconstruction coefficient. Finally, experimental results showed that the proposed method outperformed state-of-the-art algorithms. In the future, we will extend the proposed method into the scenario of multi-label image classification.

6. Acknowledgements

The National Nature Science Foundation (NSF) of China under grants 61170131, 61263035, and 61363009; the China 863 Program under grant 2012AA011005; the China 973 Program under grant 2013CB329404; the Guangxi Natural Science Foundation under grant 2012GXNSFGA060004; the funding of Guangxi 100 Plan and the Guangxi “Bagui” Teams for Innovation and Research.

References

[1] Argyriou, A., Evgeniou, T., Pontil, M., 2008. Convex multi-task feature learning. Machine Learning 73 (3), 243–272.

[2] Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: COLT. pp. 92–100.

[3] Boureau, Y.-L., Roux, N. L., Bach, F., Ponce, J., LeCun, Y., 2011. Ask the locals: multi-way local pooling for image recognition. In: ICCV. pp. 2651–2658.

[4] Cai, X., Nie, F., Huang, H., Kamangar, F., 2011. Heterogeneous image feature integration via multi-modal spectral clustering. In: CVPR. pp. 1977–1984.

[5] Chaudhuri, K., Kakade, S. M., Livescu, K., Sridharan, K., 2009. Multi-view clustering via canonical correlation analysis. In: ICML. pp. 129–136.

[6] Chen, N., Zhu, J., Xing, E., 2010. Predictive subspace learning for multi-view data: A large margin approach. Vol. 23. pp. 129–136.

[7] Dhillon, P. S., Foster, D., Ungar, L., 2011. Multi-view learning of word embeddings via cca. In: NIPS. pp. 9–16.

[8] Dhillon, P. S., Foster, D. P., Ungar, L. H., 2011. Minimum description length penalization for group and multi-task sparse learning. Journal of Machine Learning Research 12, 525–564.

[9] Efron, B., Hastie, T., Johnstone, L., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–499.

[10] Gao, S., Tsang, I. W.-H., Chia, L.-T., 2010. Kernel sparse representation for image classification and face recognition. In: ECCV. pp. 1–14.

[11] Geng, B., Tao, D., Xu, C., 2010. Daml: Domain adaptation metric learning. IEEE Transactions on Image Processing (99), 1–1.

[12] He, J., Lawrence, R., 2011. A graphbased framework for multi-task multi-view learning. In: ICML. pp. 25–32.

[13] Hou, C., Nie, F., Yi, D., Wu, Y., 2011. Feature selection via joint embedding learning and sparse regression. In: IJCAI. pp. 1324–1329.

[14] Jenatton, R., Audibert, J.-Y., Bach, F., 2011. Structured variable selection with sparsity-inducing norms. Journal of Machine Learning Research 12, 27–77.

[15] Kim, J., Monteiro, R., Park, H., 2012. Group sparsity in nonnegative matrix factorization. pp. 69–76.

[16] Kumar, A., DauméIII, H., 2011. A co-training approach for multi-view spectral clustering. In: ICML. pp. 393–400.

[17] Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2010. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research 11, 19–60.

[18] Nie, F., Huang, H., Cai, X., Ding, C., 2010. Efficient and robust feature selection via joint l2,1-norms minimization. In: NIPS. pp. 1813–1821.

[19] Nilsback, M.-E., Zisserman, A., 2006. A visual vocabulary for flower classification. In: CVPR. pp. 1447–1454.

[20] Olshausen, B. A., Field, D. J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.

[21] Owens, T., Saenko, K., Chakrabarti, A., Xiong, Y., Zickler, T., Darrell, T., 2011. Learning object color models from multi-view constraints. In: CVPR. pp. 169–176.

[22] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., Wang, P., 2010. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. The Annals of Applied Statistics 4 (1), 53–77.

[23] Quadrianto, N., Lampert, C. H., 2011. Learning multi-view neighborhood preserving projections. In: ICML. pp. 425–432.

[24] Quattoni, A., Carreras, X., Collins, M., Darrell, T., 2009. An efficient projection for l 1,∞ regularization. In: ICML. pp. 857–864.

[25] Rakotomamonjy, A., Bach, F. R., Canu, S., Grandvalet, Y., 2008. SimpleMKL. Journal of Machine Learning Research 9, 2491–2521.

[26] Saha, A., Rai, P., III, H. D., Venkatasubramanian, S., 2011. Online learning of multiple tasks and their relationships. Vol. 15. pp. 643–651.

[27] Sprechmann, P., Ramírez, I., andYonina C. Eldar, G. S., 2011. C-hilasso: A collaborative hierarchical sparse modeling framework. IEEE Transactions on Signal Processing 59 (9), 4183–4198.

[28] Sun, L., Liu, J., Chen, J., Ye, J., 2009. Efficient recovery of jointly sparse vectors. In: NIPS. pp. 1812–1820.

[29] Tan, M., Wang, L., Tsang, I. W., 2010. Learning sparse svm for feature selection on very high dimensional datasets. In: ICML. pp. 1047–1054.

[30] Wang, H., Nie, F., Huang, H., Ding, C., 2013. Feature selection via joint embedding learning and sparse regression. In: CVPR. pp. 3097–3012.

[31] Wang, H., Nie, F., Huang, H., Risacher, S. L., Ding, C., Saykin, A. J., Shen, L., ADNI, 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: ICCV. pp. 2029–2034.

[32] Wu, J., Rehg, J. M., 2011. Centrist: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8), 1489–1501.

[33] Xia, T., Tao, D., Mei, T., Zhang, Y., 2010. Multiview spectral embedding. IEEE Transactions on Systems, Man, and Cybernetics,Part B: Cybernetics 40 (6), 1438–1446.

[34] Yang, H., King, I., Lyu, M. R., 2010. Online learning for multi-task feature selection. In: CIKM. pp. 1693–1696.

[35] Yu, S., Krishnapuram, B., Rosales, R., Rao, R. B., 2011. Bayesian co-training. The Journal of Machine Learning Research 999888, 2649–2680.

[36] Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.

[37] Yuan, X., Yan, S., 2010. Visual classification with multi-task joint sparse representation. In: CVPR. pp. 3493–3500.

[38] Zhang, D., He, J., Liu, Y., Si, L., Lawrence, R. D., 2011. Multi-view transfer learning with a large margin approach. In: KDD. pp. 13–22.

[39] Zhu, X., Huang, Z., Cui, J., Shen, H. T., 2013. Video-to-shot tag propagation by graph sparse group lasso. IEEE Transactions on Multimedia 15 (3), 633–646.

[40] Zhu, X., Huang, Z., Shen, H. T., Cheng, J., Xu, C., 2012. Dimensionality reduction by mixed kernel canonical correlation analysis. Pattern Recognition 45 (8), 3003–3016.

[41] Zhu, X., Huang, Z., Wu, X., 2013. Multi-view visual classification via a mixed-norm regularizer. In: PAKDD (1). pp. 520–531.

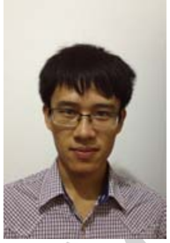
[42] Zhu, X., Huang, Z., Yang, Y., Shen, H. T., Xu, C., Luo, J., 2013. Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition 46 (1), 215–229.

[43] Zhu, X., Zhang, J., Zhang, S., 2013. Mixed-norm regression for visual classification. In: ADMA (1). pp. 265–276.

[44] Zou, H., Hastie, T., Tibshirani., R., 2006. Sparse principal component analysis. Journal of computational and graphical statistics 15 (2), 265–286.

Xiaofeng Zhu is a full professor at Guangxi Normal university, P. R. China and received his PhD degree in computer science from The University of Queensland, Australia. His research interests include large scale multimedia retrieval, feature selection, sparse learning, data preprocess, and medical image analysis. He is a guest editor of Neurocomputing, and served as a technical program committee member of several international conferences and a reviewer of over 10 international journals.

Qing Xie is a Postdoctoral Fellow in the Division of Computer, Electrical and Mathematical Sciences and Engineering (CEMSE), King Abdullah University of Science and Technology (KAUST). His research interests include Data mining, query optimization and multimedia.



Yonghua Zhu is a undergraduate student at Guangxi University, China. His research interests include data mining and machine learning.
Xingyi Liu is an associate professor at Qinzhou Institute of Socialism, Qinzhou, Guangxi, China. His research interests include data mining and pattern recognition.

Shichao Zhang received the PhD degree in computer science at the Deakin University, Australia. He is currently a China 1000-Plan distinguished professor with the Department of Computer Science, Zhejiang Gongshang University, China. His research interests include data quality and pattern discovery. He has published more than 60 international journal papers and 70 international conference papers. As a Chief Investigator, he has won 4 Australian Large ARC, 3 China 863 Programs, 2 China 973 Programs, and 5 NSFs of China. He served/is serving as an associate editor for the IEEE Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, and the IEEE Intelligent Informatics Bulletin. He also served as a PC Chair or Conference Chair for 6 international conferences. He is a senior member of the IEEE and a member of the ACM.